3-3-2014

Examining the Impact of Contextual Ambiguity on Search Advertising Keyword Performance: A Topic Model Approach

Vibhanshu Abhishek  
*Carnegie Mellon University, vibs@andrew.cmu.edu*

Jing Gong  
*Carnegie Mellon University*

Beibei Li  
*Carnegie Mellon University, beibeili@andrew.cmu.edu*

Follow this and additional works at: [http://repository.cmu.edu/heinzworks](http://repository.cmu.edu/heinzworks)

Part of the [Databases and Information Systems Commons](http://repository.cmu.edu/heinzworks) and the [Public Policy Commons](http://repository.cmu.edu/heinzworks)

---

This Working Paper is brought to you for free and open access by the Heinz College at Research Showcase @ CMU. It has been accepted for inclusion in Heinz College Research by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Examining the Impact of Contextual Ambiguity on Search Advertising 
Keyword Performance: A Topic Model Approach

Vibhanshu Abhishek, Jing Gong, Beibei Li\textsuperscript{1}

Heinz College, Carnegie Mellon University, Pittsburgh, PA - 15213

\{vibs, jingg, beibeili\}@andrew.cmu.edu

Abstract

In this paper, we explore how the contextual ambiguity of a search can affect a keyword’s performance. The context of consumer search is often unobserved and the prediction of it can be nontrivial. Consumer search contexts may vary even when consumers are searching for the same keyword. Due to the ambiguity of a keyword, a large portion of the ads displayed may fall outside a particular consumer’s interest, potentially leading to low click-through rates on search engines. In our study, we propose an automatic way of examining keyword contextual ambiguity based on probabilistic topic models from machine learning and computational linguistics. We quantify the effect of contextual ambiguity on keyword click-through performance using a hierarchical Bayesian model that allows for topic-specific effect and nonlinear position effect. We validate our study using a novel dataset from a major search engine that contains information on consumer click activities for 12,790 distinct keywords across multiple product categories from over 4.6 million impressions. We find that consumer click behaviors vary significant across keywords, and keyword category and the contextual ambiguity of the keywords significantly affect such variation. Specifically, higher contextual ambiguity can lead to a higher click-through rate (CTR) on top-positioned ads, but the CTR tends to decay faster with position. Therefore, the overall effect of contextual ambiguity on CTR varies across positions. Our study has the potential to help advertisers design keyword portfolios and bidding strategy by extracting contextual ambiguity and other semantic characteristics of keywords based on large-scale analytics from unstructured data. It can also help search engines improve the quality of displayed ads in response to a consumer search query.

Keywords: Sponsored search advertising, topic models, contextual ambiguity, machine learning.

\textsuperscript{1}Author names are in alphabetic order.
1 Introduction

With the growing pervasiveness of consumer search for relevant information and products via search engines, sponsored search advertising (also known as “paid search” or “search advertising”) has become an important marketing channel for businesses. In the first half of 2013, sponsored-search advertising generated revenues of $8.7 billion and accounted for 43% of online advertising (Interactive Advertising Bureau 2013). Most online advertising techniques offer a more effective way of targeting customers as compared to traditional advertising. However, search advertising considerably outperforms other forms of online advertisings such as display or social media advertising on metrics such as return on investments (ROI), click-through-rate (CTR), and conversion rates. The effectiveness of search advertising is attributed to the fact that search engines match the ads shown to a consumer with her current search intent derived from the keyword she uses.

When a consumer issues a query on a search engine using a keyword, for example, “hotels nyc,” the search engine identifies a list of advertisers who are bidding on the keyword “hotels nyc.” It subsequently presents an appropriate list of ads based on factors such as bids placed by the advertisers and their historical performance (Ghose and Yang 2009; Yang and Ghose 2010; Agarwal et al. 2011). The ability to present consumers ads tailored to their search context (as indicated by the keywords) considerably increases the likelihood that they will click on these ads. Put differently, sponsored search ads are pull-based ads – these ads are shown to consumers when they are searching for something specific, and tend to be relevant to consumers’ actions. Most other forms of advertising (online and offline) are push-based – these ads are shown to the consumers when they are engaged in unrelated activities, and the consumers might not have an immediate search/purchase intent when they encounter these ads. The pull-based nature of search ads leads to their relatively high performance as compared to other forms of advertising.

Even through a search keyword provides an indication of a consumer’s search context, consumers from varied contexts might use the same keyword for searching. Consumers arrive at search engines with diverse interests and their search context may vary even when they are searching using the same keyword. For example, a consumer who searches for the keyword “Mars” may be interested in astronomy and the planet Mars, or may be interested in buying chocolates and candies from the confectionery company Mars, or may be looking for a local chain of grocery stores in metropolitan Baltimore, Maryland. Similarly, a consumer who searches for “new york, new york” may be looking for tourism information about New York City, or checking about a Las Vegas hotel, or may be interested in a 1977 movie. With
advances in contextual search, search engines can predict consumers’ intent with increasing accuracy. However, the context of consumer search is not directly observed and its prediction can be nontrivial. An example of a recent search at a leading search engine in Figure 1 shows varied organic search results, which demonstrates the ambiguity the search engine faces in predicting the consumer’s search context.

Because the same keyword might refer to different contexts, competing advertisers might have different intents while bidding on a particular keyword. Firms in industries as varied as tourism, hotels, or movies might bid for the same keywords, which can lead to diverse ads as shown in Figure 2. Differences in the search context and advertisers’ intent for the same keyword make a perfect match between the two challenging (even though search advertising is often more precise as compared to other forms of advertising). In addition, due to privacy concerns and technological limitations, most search engines allow advertisers to target consumers only at the keyword level. The likelihood of an ad being clicked, *ceteris paribus*, depends on the relevance of the advertiser’s intent to the consumer’s search context. Any potential mismatch between a consumer’s search context and an advertiser’s intent might reduce the efficiency of search advertising. However, the magnitude of the mismatch varies across keywords. Some keywords are specific and do not have a variety of meaning, for example, “antivirus.” Consumers who search for “antivirus” and advertisers who bid on “antivirus” are likely to refer to the same prod-

Figure 1: Organic Search Results of the Keyword “new york, new york” on a Major Search Engine.
uct. Because keywords can have either a narrow or broad context, they might have varying appeal to consumers, and their performance might depend on the ambiguity in their context.

In this paper, we want to understand the interplay between a keyword’s context and consumers’ search behavior. More specifically, we wish to ascertain how the breadth of a keyword’s context might affect consumer behavior and keyword performance. On the one hand, prior literature in search theory suggests that as the uncertainty in the quality of search results increases, users are more likely to search (e.g., Weitzman 1979), because higher uncertainty leads to higher variance and more diversity in the alternatives, and users believe they will be more likely to find an alternative with a good fit or high value during their search. On the other hand, consumer psychology theories suggest that as the alternatives become less relevant, users are more likely to abandon their search (e.g., Jacoby et al. 1974; Dhar and Simonson 2003; Kuksov and Villas-Boas 2010), because users tend to get overwhelmed and discouraged by the complexity of information, and therefore lose their interest or trust in the search results. In reality, keyword contextual ambiguity can result in both higher diversity in ad quality and higher probability of ad irrelevancy. Therefore, how keyword contextual ambiguity would affect consumer click behavior is unclear. To explore this question, in this paper, we use a rich dataset from a major search engine to perform a cross-category analysis and examine which of these two opposing effects dominates in the context of search advertising.

One major advantage of our study is that we are able to examine a large variety of keywords across multiple product categories. Such cross-category analysis allows us to distinguish our study from the
previous research, which mostly focused on a single product category from a particular retailer (e.g., Ghose and Yang 2009; Agarwal et al. 2011; Rutz and Bucklin 2011; Yao and Mela 2011). This cross-category analysis helps us provide broad generalizations as well as focus on category specific effects. Because the effectiveness of sponsored search advertising may vary across categories, a firm may take these industry-specific insights into account while crafting its search advertising strategy. Another advantage of this dataset is the presence of all competing ads for an impression, which helps us build a richer model of consumer search and derive new insights.

Web-scale data have commonly been used in computer science research, but there is an increasing need for such analysis in marketing research and practice, as exemplified by Archak et al. (2011), Netzer et al. (2012), and Ghose et al. (2012). Due to the extensive nature of our data, we resort to novel machine learning techniques to extract textual characteristics such as keyword ambiguity, semantic category, and transactional intent. We propose an automatic method of categorizing keywords and examining keyword contextual ambiguity based on topic models from machine learning and computational linguistics. Specifically, we construct a new semantic characteristic of a keyword, topic entropy, which is derived from the results of a topic model and measures the uncertainty in predicting consumers' buying interest. Subsequently, we quantify the effect of contextual ambiguity on keyword click-through performance using a hierarchical Bayesian model.

Our analysis indicates that the semantic characteristic of keywords presented in this paper can significantly improve the prediction of keyword performance in addition to keyword characteristics that previous research (Ghose and Yang 2009; Yang and Ghose 2010; Agarwal et al. 2011; Rutz and Trusov 2011; Rutz et al. 2011) has investigated, such as keyword length, brand and location information, and keyword popularity. More specifically, we find that contextual ambiguity has two opposing effects on keyword click performance. On the one hand, higher contextual ambiguity is associated with higher baseline CTR (i.e., click probability at the top position). On the other hand, higher contextual ambiguity is also associated with a faster decay in CTR with screen position (i.e., from top to bottom). Therefore, the overall effect of contextual ambiguity on CTR for ads at various positions is a combination of these opposing effects. In other words, contextual ambiguity seems to hurt lower-positioned ads but benefit top-positioned ads.

Moreover, we also find significant interplay between keyword category and screen position. In particular, the distribution of CTR among different screen positions varies significantly across keyword categories. For example, the CTR of keywords in certain categories, such as “adult,” “home,” or “style,” is more evenly distributed across different positions compared to other categories, such as “baby products,”
“finance,” and “travel.” More specifically, we find search terms related to the latter categories tend to start out with relatively high CTR on top positions but the CTR diminishes dramatically as position goes down. However, search terms related to the former categories tend to start out with fewer clicks on top positions, but continue to receive clicks on ads at lower positions. In fact, we find that the click depth (i.e., lowest position on the screen) for adult-related ads is the greatest among all the categories considered in our analysis. These results suggest that position effects appear to be more significant for certain product categories than others. It is critical for advertisers and search engines to understand the heterogeneity in keyword performance across different categories.

This paper makes the following contributions. First, we demonstrate how machine learning and computational linguistic tools such as topic models can be used to extract keyword categories, contextual ambiguity, and other semantic characteristics of keywords based on large-scale analytics from unstructured text data. Using a topic model, we are able to introduce a new keyword attribute – topic entropy – that quantifies the contextual ambiguity of a keyword, which is critical in understanding keyword performance. Second, we expand the sponsored search advertising literature by examining how keyword attributes affect keyword performance (i.e., CTR) in multi-category sponsored search advertising. Most previous studies obtain data from a single advertiser and thus can only measure keyword performance from the perspective of one specific firm, and the results may lack generalizability. The availability of detailed consumer search data for keywords across multiple advertisers and categories allows us to measure click performance at position level and identify keywords that have the potential to generate clicks – even at lower ad positions. Third, we find that consumer click behaviors vary significantly across keywords, and such variation can be explained by keyword category and the contextual ambiguity of keywords, even after controlling for keyword length, popularity, and other observed characteristics. Our empirical analysis offers new insights into consumer search behavior in the context of multi-category sponsored search, which can help advertisers select and evaluate keywords more effectively. For instance, firms may want to better allocate their campaign resources by advertising on less ambiguous keywords or bidding for higher positions on ambiguous keywords. Our findings can also help search engines understand the manner in which consumers respond to sponsored advertising, and design better advertising platforms. Interestingly, our analysis suggests that consumers follow a two-step process while clicking on search ads. The effect of different keyword characteristics, such as entropy, number of words, and the presence of a brand name, differs in the propensity to start clicking and traversing the list of search ads. As a result, we reconcile the two opposing theories on consumer search and paves the path for richer consumer search models that can incorporate our findings.
The remainder of this paper is organized as follows. Section 2 outlines related streams of literature. Section 3 describes the click-through data used in our empirical study. Section 4 presents our topic modeling approach to categorize keywords and quantify keyword contextual ambiguity and our approach to extract additional keyword semantic features. Section 5 describes how we incorporate keyword characteristics and non-linear position effects in a hierarchical Bayesian framework. We present the main results in section 6 and the robustness checks in section 7. We conclude with a discussion in section 8.

2 Literature Review

Our paper is closely related to three different streams of literature – (i) search advertising, (ii) machine learning and text mining, and (iii) search theory.

First, our study is related to the sponsored search advertising literature. During the past decade, the increasing popularity of sponsored search advertising has motivated research from multiple areas. Theoretical analysis on sponsored search include Edelman et al. (2007), Varian (2007), Athey and Ellison (2011), Katona and Sarvary (2010), and so on. Most of these studies focus on auction design and bidding strategies of firms. Empirical research on search advertising is also growing rapidly (Ghose and Yang (2009); Yang and Ghose (2010); Agarwal et al. (2011); Animesh et al. (2011); Rutz et al. (2011)). Most of these studies primarily use keyword-level aggregate data provided by advertisers in particular industries to study ad performances. Exceptions are Jerath et al. 2013, Jeziorski and Segal 2012, and Yao and Mela 2011, who use individual-level data provided by search engines. Appendix A provides an overview of previous empirical studies on sponsored search advertising. The keyword characteristics the previous studies examined include keyword length, brand and retailer name (e.g., Ghose and Yang 2009; Yang and Ghose 2010; Jerath et al. 2013), geographic location (Rutz et al. 2012), specificity (Agarwal et al. 2011; Jerath et al. 2013), and product categories (Animesh et al. 2010). However, in these papers, keyword characteristics are handpicked on a small scale. Rutz et al. (2011) is to our best knowledge the most relevant to our study. The authors use a top-down approach to identify the key area of business related to each keyword (“keyword cluster”) in the automobile industry. However, this process relies on human input (e.g., interviews, questionnaires, and/or other communications with the firm’s management) to define keyword clusters. In our study, the use of topic modeling, namely, Latent Dirichlet Allocation model (LDA; Blei et al. 2003), allows us to characterize the topical content and category of a keyword automatically on a large scale using unstructured text data. Moreover, most of the aforementioned
studies focus on only a small set of keywords using data from a particular advertiser. In our study, we use a dataset that contains consumer click-through information for a large number of keywords across multiple product categories from different types of advertisers. By analyzing the contextual information extracted from each keyword, we are able to examine keyword performance across multiple product categories under different search contexts.

Second, our study is also related to literature in machine learning and text mining in marketing. A stream of research has recently emerged in marketing and information systems that applies machine learning and text mining techniques in examining online markets (e.g., Archak et al. 2011; Netzer et al. 2012; Ghose et al. 2012). Ghose and Ipeirotis (2011) and Archak et al. (2011) use text mining to extract multiple aspects of online review texts to identify important text-based features and to study their impact on review helpfulness (Ghose and Ipeirotis 2011) and product sales (Archak et al. 2011). Netzer et al. (2012) combine text mining and semantic network analysis to understand the brand associative network and the implied market structure. Decker and Trusov (2010) use text mining to estimate the relative effects of product attributes and brand names on the overall evaluation of the products. Ghose et al. (2012) use text mining and image classification to analyze the economic effects of user-generated content and crowd-sourced content, and to design a novel utility-based ranking system for products that can lead to an increase in consumer surplus. In our study, we propose applying topic modeling (i.e., LDA; Blei et al. 2003) from machine learning and natural language processing that allows us to extract the topical content of each keyword. To the best of our knowledge, our work is the first study that introduces probabilistic topic modeling in marketing research.

Third, our research is related to research on consumer search behavior. Traditional economic theories model consumer search behavior in an expected utility framework, where consumers search sequentially and stop searching when the marginal cost of search outweighs the marginal gain from search (Weitzman 1979). Many recent empirical studies adopt the sequential search framework to analyze online consumer search behavior (e.g., Kim et al. 2010; Ghose et al. 2013). This literature concludes that as the uncertainty in the quality of the alternatives becomes higher, consumers are more likely to search. On the other hand, work by Iyengar and Lepper (2000), Dhar and Simonson (2003), and Kuksov and Villas-Boas (2010) suggests that as the number of alternatives increases, consumers are less likely to search. Too many choices may present a cognitive burden on the consumers, reducing the likelihood of a purchase. Our study provides a rich and large dataset to shed light on the manner in which consumers perform search activities in a search advertising environment and reconcile these opposing viewpoints.
3 Data

The dataset used in this study is provided by one of the largest search engines in the United States. It consists of a random sample of close to 8 million search impressions conducted in the United States between August 10, 2007 and September 16, 2007. For every impression, the dataset comprises the keyword a user searches and a list of search ads shown to the consumer. The maximum number of ads shown per impression is eight. Although we do not have individual identifiers, which may restrict our ability to track individuals over time, every ad in our dataset has a unique identifier. For each ad displayed, we observe whether it was clicked during an impression. Note that although we have a unique ad identifier and we can track an ad across impressions, we do not have any ad-specific information. Because our data are derived from the search engine, we do not have post-click information, unlike some previous papers (Ghose and Yang 2009; Agarwal et al. 2011), which use data provided by advertisers.

We apply the following steps to pre-process the data: (i) we focus on keywords that receive at least one click during the entire five-week period and (ii) remove keywords that are domain names. We remove the low-performing keywords for two specific reasons. First, these keywords are not as relevant, as ads related to them never get clicked in our sampling period. Second, applying our machine learning techniques to all the keywords would take a substantial amount of time. Hence, we restrict our analysis to a smaller but more relevant dataset. We choose to ignore keywords containing domain names, because users who use these keywords know exactly which website they wish to visit, and these keywords are unlikely to lead to additional traffic for the website.

The subset of the dataset we examine in our analysis includes 12,790 distinct keywords from more than 4.6 million impressions. More than 0.17 million unique ads are displayed when consumers search for these keywords, resulting in 5.19 ads per impression. Twelve percent of the impressions receive at least one click. Table 1 presents the distribution of the number of clicks. Overall, there are about 640,000 clicks, and the average number of clicks per impression is 0.14. This observation is in agreement with prior research by Jerath et al. (2013) that suggests very few searches lead to clicks on search ads because the user needs might be met by organic ads.

Table 1 presents summary statistics for the data. Note that we use the impression-level data to generate the summary statistics at a keyword level. IMP denotes the total number of times consumers search for a particular keyword in the dataset, which is analogous to the popularity measure used in Jerath et al. (2013). CLICKS measures the total number of clicks a particular keyword receives. CLICK_DEPTH measures the average number of clicks per impression of a keyword. AVG_CTR
Table 1: Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP</td>
<td>362.919</td>
<td>3736.077</td>
<td>2</td>
<td>292,692</td>
</tr>
<tr>
<td>CLICKS</td>
<td>49.660</td>
<td>197.835</td>
<td>1</td>
<td>6,984</td>
</tr>
<tr>
<td>CLICKS DEPTH</td>
<td>0.186</td>
<td>0.167</td>
<td>0</td>
<td>1.522</td>
</tr>
<tr>
<td>AVG_CTR</td>
<td>0.028</td>
<td>0.027</td>
<td>0</td>
<td>0.555</td>
</tr>
<tr>
<td>NUM_WORDS</td>
<td>1.741</td>
<td>0.683</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>AVG_NUM_ADS</td>
<td>6.599</td>
<td>1.215</td>
<td>1.139</td>
<td>8</td>
</tr>
<tr>
<td>AVG_AD_QUALITY</td>
<td>0.026</td>
<td>0.022</td>
<td>0.001</td>
<td>0.451</td>
</tr>
</tbody>
</table>

measures the average CTR of ads across different impressions for a keyword. NUM_WORDS denotes the number of words in the keyword. Previous studies on sponsored search report inconsistent findings regarding the effect of keyword length on CTR. Ghose and Yang (2009) suggest a negative relationship between keyword length and CTR, whereas Rutz et al. (2012) find that as the keyword becomes longer, the CTR increases. In this study, we try to resolve this conflict. AVG_NUM_ADS measures the average number of competing advertisers during an impression, which denotes the competitive intensity for a keyword. An important factor that determines keyword performance is the quality of ads, which is captured in AVG_AD_QUALITY. Because we do not directly observe ad attributes, we proxy for the quality of ads associated with a particular keyword by the average CTR that these ads receive for other keywords. Search engines commonly use this measure to predict ad quality. In section 4, we discuss how to use machine learning tools to extract additional keyword semantic features, such as contextual ambiguity, brand, location, and transactional intent. Appendix B reports the correlation matrix among variables.

The unique large dataset we use allows us to extend the search advertising literature and provide insights that can be generalized across multiple categories. This cross-category analysis can also help us identify differences between keywords across different categories. Appendix A presents a comparison between the data used in our paper and prior research in search advertising is presented.

4 Extracting Semantic Features Using Machine Learning

In this section, we describe the machine learning techniques used to extract keyword characteristics. Given the large number of keywords used in our analysis, we resort to unsupervised machine learning
methods. Incidentally, these methods work fairly well when applied to large datasets as in our analysis. We first describe a topic model that is used to categorize the keywords used in this study. We subsequently use this categorization to compute keyword contextual ambiguity. In addition, we also propose automated methods to extract keyword characteristics such as the presence of a brand name and the transactional nature of a keyword.

4.1 A Generative Model of the Topical Content of Keywords

The major challenge in examining the impact of a keyword’s contextual ambiguity on consumer click behavior is how to quantify such ambiguity. We model the contextual ambiguity of each keyword based on probabilistic topic models from machine learning and natural language processing. Topic models are unsupervised algorithms that aim to extract hidden topics from unstructured text data. The intuition behind topic models is that a topic is a cluster of words that frequently occur together, and that documents, consisting of words, may belong to multiple topics with different probabilities. A probabilistic topic model tries to discover the underlying topic structure in a statistical framework. Topic models have been applied to many research areas, including the analysis of social network textual data (e.g., Paul and Dredze 2011), opinion mining, and information retrieval (e.g., Titov and McDonald 2008; Aral et al. 2011).

4.1.1 Corpus Construction and Data Pre-processing

We first construct a corpus of documents that describe the information content conveyed by the keywords. Keywords are usually words or short phrases. Obtaining adequate semantic information based on only keywords is usually difficult. To solve this challenge, we use Google organic search results to augment the keyword data and better understand the semantic meaning of each keyword (Dai et al. 2006; Abhishek and Hosanagar 2007). Obtaining the true contextual meaning(s) of a keyword can be difficult, but Google organic search results generated based on the classical theory of document relevancy provide a reasonable approximation.

Therefore, for all 12,790 keywords in our dataset, we extract the title and textual content of the brief description from each of the top-50-ranked Google organic search results (Figure 1), to construct the corresponding keyword-specific document. The results produce a total of 12,790 documents, each containing the most relevant information describing the corresponding keyword. After constructing the corpus of keyword documents, we pre-process the data following a standard procedure (e.g., Aral et al. 2011). We first remove annotations and tokenize the sentence into distinct terms. Then we remove
stopping words using a standard dictionary.

4.1.2 Latent Dirichlet Allocation

We use topic models to automatically infer semantic interpretations of keyword meanings. The most widely used topic model is the Latent Dirichlet Allocation model (LDA; Blei et al. 2003), which is a hierarchical Bayesian model that describes a generative process of document creation. Previous research shows that humans tend to agree with the coherence of the topics generated by LDA, which provides strong support for the use of topic models for information retrieval applications (Chang et al. 2009).

The goal of LDA is to infer topics as latent variables from the observed distribution of words in each document. In particular, a topic is defined as a multinomial distribution over a vocabulary of words, a document is a collection of words drawn from one or more topics, and a corpus is the set of all documents. Based on the discussion in subsection 4.1.1, we construct a document for each keyword that best reflects the contextual information of the keyword. We now discuss how we use LDA to infer the topics from the corpus of documents.

Formally, let $T$ be the number of topics related to the corpus, let $D$ be the number of documents in the corpus, and let $W$ be the total number of words in the corpus. We assume that each document in the corpus is generated according to the following process:

Step 1. For each topic $t$, choose $\phi_t = (\phi_{t1}, ..., \phi_{tW}) \sim Dirichlet(\psi)$, where $\phi_t$ describes the word distribution of topic $t$ over the vocabulary of words.

Step 2. For each document $d$, choose $\theta_d = (\theta_{d1}, ..., \theta_{dT}) \sim Dirichlet(\omega)$, where $\theta_{dt}$ is the probability of topic $t$ to which document $d$ belongs.

Step 3. For each word $n$ in document $d$, (1) choose a topic $t_{dn} \sim Multinomial(\theta_d)$, and (2) choose a word $w_{dn} \sim Multinomial(\phi_{t_{dn}})$.

$\psi$ and $\omega$ are hyper-parameters for the two prior distributions - $Dirichlet(\psi)$ as the prior distribution of $\phi$ (word distribution in a topic) and $Dirichlet(\omega)$ as the prior distribution of $\theta$ (topic distribution in a document). We use the values suggested by Steyvers and Griffiths (2007) ($\psi = 0.01$ and $\omega = 50/T$).

Based on the generative process described above, we use a Markov chain Monte Carlo (MCMC) algorithm to estimate $\phi$ and $\theta$. Specifically, we use a collapsed Gibbs sampler to sequentially sample the topic of each word token in the corpus conditional on the current topic assignments of all other word tokens (see Griffiths and Steyvers 2004 for details). We run a collapsed Gibbs sampler using MALLET (McCallum 2002) with 2,000 iterations. For each keyword, we obtain the posterior topic probabilities inferred from its corresponding document of Google organic search results. In our study, we estimate
the LDA model with a different number of topics, $T = 20, 50,$ and $100$.

The most frequent words identified for the 20-topic model are presented in Figure 3, where topics are color coded. For convenience, we assign a label to each topic (e.g., “Sport,” “Music,” “Food”) based on its high-frequency words. For example, documents related to “style” often contain words such as “dress,” “party,” “woman,” and “fashion,” and so on. In addition, words that occur frequently in multiple topics are highlighted in brown, such as “free,” “shop,” and “find.”

![Figure 3: Frequent Words in Each Topic](image)

Figure 4 illustrates keywords’ topic distributions estimated from the 20-topic model for a sample of keywords. In Figure 4, topics are labeled on the horizontal axis, and keywords are labeled on the vertical axis. Each bubble indicates a posterior topic probability, with larger bubbles representing higher probabilities. For example, the top-left bubble represents the posterior probability that the keyword “judges gavels” belongs to the topic “music,” which is much smaller than the posterior probability that
“judges gavels” belongs to “government,” represented by the eighth bubble on the first row. Meanwhile, the keyword “marriage records” has a much larger posterior probability of belonging to the topic “government,” which suggests “marriage records” is most likely related to government affairs rather than other topics.

![Figure 4: Topic Distribution of Sample Keywords](image)

### 4.1.3 Entropy as a Measure for Keyword Ambiguity

We propose using *topic entropy* to measure keyword ambiguity. It captures the uncertainty of a keyword/document's topic distribution (Hall et al. 2008). In information theory, entropy measures the unpredictability of a random variable. In our context, each keyword is associated with its own topic
distribution inferred from the keyword-specific document. Therefore, we treat the topic assignment as a multinomial random variable, and use topic entropy to quantify how "noisy" a keyword is in terms of underlying topics. The higher the entropy is, the more ambiguous the keyword. In other words, keywords with higher entropy tend to relate to a broader range of topics (more ambiguous), whereas keywords with lower entropy tend to relate to fewer dominant topics (less ambiguous).

More formally, let $\hat{\theta}_{kt}$ denote the posterior probability that keyword $k$ belongs to topic $t$. We therefore define the topic entropy of keyword $k$ as follows:

$$\text{TOPIC\_ENTROPY}_k = -\sum_{t=1}^{T} \hat{\theta}_{kt} \log(\hat{\theta}_{kt}),$$

where $T$ is the total number of topics.

![Figure 5: Topic Entropy: A Demostration](image)

Figure 5: Topic Entropy: A Demostration
Figure 5 illustrates the posterior topic probabilities and computed entropy for two keywords, “free anti virus” and “express.” For the keyword “free anti virus,” the estimated probability that it is related to topic “computer” is extremely high (0.93), and low for other topics, which means “free anti virus” is highly likely to relate to a single dominant topic – “Computer.” As a result, the computed topic entropy for “free anti virus” is relatively small (0.44). By contrast, the keyword “express” has a fairly even topic distribution, resulting in relatively high topic entropy (2.64). Consequently, predicting what consumers are looking for when they search for the keyword “express” is difficult. We present the summary statistics for the estimated topic entropy in Table 2. As can be seen in Table 2, the maximum entropy value depends on the number of topics chosen. Simple calculation also shows that with $T$ topics, entropy ranges from 0 to $\log(T)$.\(^2\) The high correlations among entropy values derived based on a different number of topics also suggest entropy seems to be fairly robust to the number of topics specified in the LDA model.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20 Topics</td>
</tr>
<tr>
<td>20 Topics</td>
<td>1.602</td>
<td>0.451</td>
<td>0.338</td>
<td>2.990</td>
<td></td>
</tr>
<tr>
<td>50 Topics</td>
<td>2.020</td>
<td>0.518</td>
<td>0.426</td>
<td>3.898</td>
<td>0.884</td>
</tr>
<tr>
<td>100 Topics</td>
<td>2.290</td>
<td>0.554</td>
<td>0.489</td>
<td>4.581</td>
<td>0.862</td>
</tr>
</tbody>
</table>

4.2 Extracting Other Keyword-level Semantic Features

Previous studies use brand name (e.g., Ghose and Yang 2009; Yang and Ghose 2010; Jerath et al. 2013) and location information (Rutz et al. 2012) to predict keyword performance. Extraction of such keyword characteristics usually relies on a human expert to determine whether a keyword contains brand or location information. With a large number of keywords, manual classifications become costly. Therefore, we apply an automatic way of extracting brand and location information based on “fuzzy” string matching (Elmagarmid et al. 2007). We describe the detailed procedures in subsection 4.2.1.

In addition to brand and location information, several studies also suggest that understanding consumers’ search intent is important (e.g., Moe 2003; Dai et al. 2006). Consumers may have different search intent, and the keywords they use in their search activities may reflect their intent. Broder

\(^2\) The number of topics $T$ is pre-specified before estimating the LDA model. Entropy for keyword $k$ is the smallest when there exists $t \in \{1, \ldots, T\}$ such that $\hat{\theta}_{kt} = 1$; Entropy is the largest when for all $t \in \{1, \ldots, T\}$, $\hat{\theta}_{kt} = 1/T$. 

16
(2002) proposes three types of search goals: navigational, informational, and transactional. In this study, we are interested in learning how likely consumers are to engage in a transaction when they search for a keyword. Therefore, we focus on detecting transactional intent from keywords.

4.2.1 Extracting Brand and Location Information

We used a rule-based method to identify whether a keyword contains brand information. First, we obtain a list of brand names from namedevelopment.com, and use a fuzzy string matching algorithm to match each keyword against the list of brand names. In particular, we use Levenshtein distance (also called edit distance; Levenshtein 1966) to measure string similarity.\(^3\) Using partial matching, we allow substrings of a keyword to match against brand names. For example, we want to match the keyword “ikea store” to the brand “ikea.” For each keyword, we identify the brand name that gives the longest partial string match. We classify the keyword as containing brand information if one of the following conditions is met: (1) if the highest full-string similarity is greater than 0.85 (i.e., Levenshtein distance computed from our model); (2) if the highest partial-string similarity is greater than 0.85, and the brand name is a complete word in the keyword other than a substring of a word. We choose a Levenshtein distance of 0.85 as the cut-off point to allow for a moderate level of mis-spelling. For example, we match the keyword “chipolte” with the brand “chipotle,” and “walmart” with “wal mart.” \(^4\)

We use a similar approach to extract whether a keyword contains location information. We obtain a list of US city and state names, and match each keyword against the list of locations. For each keyword, we find the location name that gives the longest string match. We classify the keyword as containing location information if the highest partial string similarity is 1, which means an exact match is found, and the location name is a complete word in the keyword.\(^5\)

4.2.2 Extracting Transactional Intent

Some keywords may contain explicit transactional words, such as “cheap hotels” and “cruise deals,” but most keywords don’t contain explicit transactional indicators in the keywords, such as “airline tickets” and “honda parts.” The augmented Google organic search results, on the other hand, provide a better picture in terms of consumer search intent. If the keyword has a transactional intent, the Google organic

\(^3\)As a robustness check, we also use the “n-grams” method for string matching, where we define n=2, 3, and 4. We find the final results remain consistent.

\(^4\)To choose the optimal cut-off distance, we first manually identified whether a smaller number of keywords contain brand names. We then tried different cut-off values and chose the one (e.g. 0.85) that minimizes classification errors on the small keyword set.

\(^5\)Similar to the process of identifying brand names, we chose the optimal cut-off distance based on a smaller set of keywords to minimize the classification error.
search results are likely to contain transactional indicators such as “buy,” “discount,” “promotion,” and “check out.” Therefore, we propose to infer transactional intent using the keyword’s corresponding Google organic results. First, we compose a list of transactional words based on Dai et al. (2006) and general knowledge. The terms are listed in Table 3. Then, for each keyword, we count the frequency of transactional words in the corresponding Google organic results. We use LOG_TRANS, the natural log of the frequency of transactional words, to measure keyword’s transactional intent.

Table 3: Transactional Words

<table>
<thead>
<tr>
<th>advertise</th>
<th>brand</th>
<th>cost</th>
<th>get</th>
<th>price</th>
<th>rent</th>
<th>service</th>
<th>auction</th>
<th>cart</th>
<th>coupon</th>
<th>gift</th>
<th>promo</th>
<th>reserve</th>
<th>ship</th>
</tr>
</thead>
<tbody>
<tr>
<td>bidding</td>
<td>cheap</td>
<td>customer</td>
<td>lease</td>
<td>promotion</td>
<td>retail</td>
<td>shop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bill</td>
<td>check out</td>
<td>deal</td>
<td>market</td>
<td>product</td>
<td>sale</td>
<td>store</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>clearance</td>
<td>delivery</td>
<td>offer</td>
<td>purchase</td>
<td>saving</td>
<td>ticket</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>buy</td>
<td>consumer</td>
<td>discount</td>
<td>pay</td>
<td>rebate</td>
<td>sell</td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2.3 Summary Statistics of Other Extracted Features

We present summary statistics of other extracted semantic features in Table 4. BRAND is a dummy variable indicating whether the keyword string contains brand names. LOCATION is a dummy variable indicating whether the keyword contains city or state names. LOG_TRANS measures a keyword’s transactional intent. In our data, 17% of the keywords are classified as containing brand information, and 13% are classified as containing location information. We present the correlation matrix among all variables in Appendix A.

Table 4: Summary of Other Extracted Features

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAND</td>
<td>0.172</td>
<td>0.378</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.130</td>
<td>0.336</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>2.173</td>
<td>1.075</td>
<td>0</td>
<td>5.509</td>
</tr>
</tbody>
</table>
5 A Model of Keyword Performance

Because the focus of this paper is on estimating the effect of keyword characteristics on CTR, we build a keyword-position model to capture how CTR varies with position. Estimating any keyword-specific model using impression-level data would take an extremely large amount of time, hence we choose to perform our analysis at a keyword-position level instead of at an impression level. We begin this discussion by formally introducing a hierarchical Bayesian model for keyword performance.

5.1 Modeling CTR

Feng et al. (2007) analyze a large dataset from a major search engine and suggest the exponential decay function provides the best fit for the variation in CTR with position. Following Feng et al. (2007) and Abhishek and Hosanagar (2013), we assume the CTR for a keyword follows an exponential decay function. More specifically, the CTR for an ad at position \( p \) for keyword \( k \) that belongs to topic \( t \) is as follows:

\[
CTR_{kpt} = P(\text{click}_{kp} = 1|\text{topic}_k = t) = \alpha_{kt} \gamma_{kt}^{p-1},
\]

where \( \alpha_{kt} \) measures the CTR at the topmost position, and \( \gamma_{kt} \) measures the relative decrease in CTR as the position increments by one.\(^6\) Both \( \alpha_{kt} \) and \( \gamma_{kt} \) are constrained to be between 0 and 1, as the CTR at the top position is always less than 1, and the CTR decreases as we move to ads lower in the list. Here \( \alpha_{kt} \) also represents the baseline CTR associated with ads for keyword \( k \); in other words, as \( \alpha_{kt} \) increases, ads at all positions are more likely to get clicked. On the other hand, when \( \gamma_{kt} \) increases, the CTR decreases slowly with position. Incidentally, a model of sequential search where consumers start out at the top of the list and traverse the list downward would give rise to an identical model. In such a model, the probability of continuing search and moving to an ad in the subsequent position is given by \( \gamma_{kt} \). These parameters differ across topics as consumers who are searching for a specific topic might demonstrate different kinds of behavior due to the nature of the product category or the information provided by organic search results. This model implicitly assumes consumer search behavior is i.i.d, which has been commonly used in the search advertising literature (Ghose and Yang 2009; Abhishek and Hosanagar 2013). To ensure \( \alpha_{kt} \) and \( \gamma_{kt} \) are constrained between 0 and 1, we use the logistic link function. Specifically, we assume

\[
\alpha_{kt} = \frac{e^{\exp(\tilde{\alpha}_{kt})}}{1 + e^{\exp(\tilde{\alpha}_{kt})}}, \quad \text{and} \quad \gamma_{kt} = \frac{e^{\exp(\tilde{\gamma}_{kt})}}{1 + e^{\exp(\tilde{\gamma}_{kt})}}.
\]

\(^6\) Here, position indicates the distance from the top of the list. Hence smaller positions are better than larger positions.
The number of clicks received by all ads at at position $p$ for keyword $k$ is modeled as a binomial distribution. The probability of observing $\text{CLICKS}_{kp}$ clicks out of $\text{IMP}_{kp}$ is given by

$$\text{Binom}(\text{CLICKS}_{kp}; \text{IMP}_{kp}, \text{CTR}_{kp}) = \Pr(\text{CLICKS}_{kp}|\text{IMP}_{kp}, \text{topic} = t) = \binom{\text{IMP}_{kp}}{\text{CLICKS}_{kp}} \text{CTR}_{kp}^{\text{CLICKS}_{kp}} (1 - \text{CTR}_{kp})^{\text{IMP}_{kp} - \text{CLICKS}_{kp}}$$

where $\text{Binom}(.; n, p)$ represents the probability mass function of a binomial distribution with $n$ trials and a probability of success equal to $p$.

### 5.2 Incorporating Heterogeneity

To capture the impact of keyword characteristics on CTR and how CTR decreases with positions, we use a hierarchical Bayesian model that allows for topic-specific effect. Hierarchical Bayesian models have been widely used in marketing (Rossi and Allenby 2003) and sponsored search literature (e.g. Ghose and Yang 2009; Yang and Ghose 2010; Agarwal et al. 2011). To incorporate heterogeneity across keywords, we assume $\bar{\alpha}_{kt}$ and $\bar{\gamma}_{kt}$ follow a multivariate normal distribution:

$$(\bar{\alpha}_{kt}, \bar{\gamma}_{kt}) \sim \text{MVN}(\mu_{kt}, \Phi),$$

where $\mu_{kt} = (\mu_{kt}^{(a)}, \mu_{kt}^{(\gamma)})'$. Note the distribution of $\bar{\alpha}_{kt}$ and $\bar{\gamma}_{kt}$ depends on the topic $t$. We further assume the mean effects $\mu_{kt}^{(a)}$ and $\mu_{kt}^{(\gamma)}$ are as follows:

$$\mu_{kt}^{(a)} = \beta_{0t}^{(a)} + X_k' \beta^{(a)} \quad \text{and} \quad \mu_{kt}^{(\gamma)} = \beta_{0t}^{(\gamma)} + X_k' \beta^{(\gamma)},$$

where $\beta_{0t}^{(a)}$ and $\beta_{0t}^{(\gamma)}$ denote the topic-specific intercept terms, and $X_k'$ is a vector of characteristics for keyword $k$, including $\text{TOPIC\_ENTROPY}$, $\text{NUM\_WORDS}$, $\text{BRAND}$, $\text{LOCATION}$, $\text{LOG\_TRANS}$, $\text{AVG\_AD\_QUALITY}$, $\text{AVG\_NUM\_AD}$ and $\text{LOG\_IMP}$. We assume the intercept terms are drawn from a multivariate normal distribution as follows:

$$(\beta_{0t}^{(a)}, \beta_{0t}^{(\gamma)}) \sim \text{MVN}(\mu_0, \Omega_0),$$
where $\mu_0 = (\mu_0^{(\alpha)}, \mu_0^{(\gamma)})'$. We assume multivariate normal priors on $\beta^{(\alpha)}$, $\beta^{(\gamma)}$, and $\mu_0$, and Inverse-Wishart priors on $\Phi$ and $\Omega_0$.

$$\beta^{(\alpha)} \sim MVN(0, \Sigma^{(\alpha)}) \quad \beta^{(\gamma)} \sim MVN(0, \Sigma^{(\gamma)})$$

$$\mu_0 \sim MVN(0, \Sigma_0)$$

$$\Phi \sim \text{InverseWishart}(V^\Phi, \nu^\Phi)$$

$$\Omega_0 \sim \text{InverseWishart}(V^\Omega, \nu^\Omega)$$

where $\Sigma^{(\alpha)} = \Sigma^{(\gamma)} = 1000I_8$, $\Sigma_0 = V^\Phi = V^\Omega = 1000I_2$, and $\nu^\Phi = \nu^\Omega = 2$.

In summary, we include heterogeneity in keyword performance at several levels: (i) heterogeneity across topics is introduced using topic-specific intercept terms, (ii) observed heterogeneity between keywords is introduced using the keyword characteristics, and (iii) unobserved heterogeneity is introduced using the multi-variate distribution in equation (4).

5.3 Likelihood Function and Topic Proportions

In the previous section, we describe a keyword’s performance conditional on the membership in a topic. The likelihood of observing the data for a keyword conditional on the topic is given by equation (3). However, as we discuss in section 4, each keyword has a distribution over the list of topics. The unconditional probability of observing the data for a particular keyword is as follows:

$$L(\text{Data} | \text{keyword} = k) = \prod_p \sum_t \hat{\theta}_{kt}.\text{Binom}(\text{CLICKS}_{kp}; \text{IMP}_{kp}, \text{CTR}_{kpt}).$$

Here, $\hat{\theta}_{kt}$ represents the probability that keyword $k$ belongs to topic $t$, as estimated in section 4. Note that this formulation is different from a latent class model, as we do not estimate $\hat{\theta}_{kt}$ simultaneously with other parameters, but derive it from the LDA model estimated in section 4. The overall log-likelihood of observing the data is given by

$$LL(\text{Data} | \{\hat{\alpha}_{kt}, \{\hat{\gamma}_{kt}\}, \beta^{(\alpha)}_0, \beta^{(\gamma)}_0, \beta^{(\alpha)}, \beta^{(\gamma)}, \mu_0, \Phi, \Omega_0\}) = \sum_k \sum_p \log \left( \sum_{k'} \hat{\theta}_{kt}.\text{Binom}(\text{CLICKS}_{kp}; \text{IMP}_{kp}, \text{CTR}_{kpt}) \right)$$

where $\beta^{(\alpha)}_0$, $\beta^{(\gamma)}_0$, $\beta^{(\alpha)}$, $\beta^{(\gamma)}$, $\mu_0$, $\Phi$, and $\Omega_0$ are the parameters to be estimated from the model.
6 Empirical Analysis and Results

The previous section outlines the model used in the empirical analysis. We estimate the hierarchical Bayesian model using Markov Chain Monte Carlo (MCMC) techniques (Gelman et al. 2003) presented below and proceed to discuss our results.

6.1 Model Estimation

We estimated the model using MCMC method (Gelman et al. 2003). In particular, we use WinBUGS, a software package that applies a Gibbs sampling approach to iteratively sample unknown parameters from their conditional distribution given the data and all other variables. (We discuss the details of the Gibbs sampling procedure in Appendix D.) We mean-centered all keyword characteristics so that the intercepts $\beta_{0t}^{(\alpha)}$ and $\beta_{0t}^{(\gamma)}$ can be interpreted as estimates of $\tilde{\alpha}_{kt}$ and $\tilde{\gamma}_{kt}$ for a typical keyword of which the covariates are set to mean values (equation 5). We use 70% of the keywords (8,953 unique keywords chosen at random) as the estimation sample. For the other 3,837 keywords, we perform out-of-sample prediction analysis in section 7. The summary statistics for both samples are reported in Appendix C. Comparing Panel A and Panel B in Appendix C shows that the two samples are fairly similar. Based on the estimation sample, we ran two MCMC chains, each with 70,000 iterations. We discarded the first 35,000 iterations of each chain, and mixed the last 35,000 iterations of the two chains to compute summary statistics of the posterior distribution of the parameters. For interpretability, we use the topic distributions and entropy estimated from a 20-topic LDA model for the HB estimation.

6.2 Estimation Results

We use the hierarchical Bayesian model presented in section 5 to explore how consumer click behavior differs across keywords with variation in contextual ambiguity. We first discuss the estimation results for the CTR on the top position, as captured by $\alpha$, and then discuss the estimation results for the decay parameter, $\gamma$.

6.2.1 Effect on $\alpha$

The estimated effect of different factors on the baseline CTR ($\alpha$) is presented in Table 5. First, the effect of topic entropy on $\alpha$ is positive and statistically significant, indicating that keywords with higher topic entropy are associated with higher overall CTR. This finding is consistent with the effect of “choice

---

7We experimented with different numbers of topics (e.g., 20, 50, and 100) in our preliminary analysis, and the results are robust to the number of topics.
uncertainty” on pre-purchase search activities as observed in behavioral studies (Urbany et al. 1989; Lanzetta 1963). A consumer who searches with a higher entropy keyword may face higher uncertainty about alternatives of a consideration set, and may be more likely to explore the ads associated with the ambiguous keyword to formulate a better understanding of alternatives. In this regard, our finding also concurs with extant literature on consumer search (e.g., Weitzman 1979; Kim et al. 2010). Classical consumer search theory suggests that as the uncertainty in the search results increases, users are more likely to search (e.g., Weitzman 1979), because higher uncertainty reflects higher variance in the distribution of the utility of the alternatives, and users believe they will be more likely to find an alternative with an extreme value in the utility distribution (e.g., good fit or high value) during their search. Therefore, the expected marginal benefit of search is likely to be high when the choice uncertainty is high, which explains the higher probability of consumer search on average. Another explanation for the positive effect of ambiguity on $\alpha$ can be provided by the organic results shown for the search. The search for ambiguous keywords leads to a varied set of results (as shown in Figure 1) by construction, which might force users to click on ads to fulfill their search intent. Incidentally, Ghose and Yang (2009) also report that longer keywords, which are less ambiguous, are associated with lower CTR. However, unlike Ghose and Yang (2009), we find that $NUM\_WORDS$ does not have a statistically significant impact on the baseline CTR after controlling for topic entropy.

Table 5: Estimation Results for $\alpha$

<table>
<thead>
<tr>
<th>Baseline Effect ($\alpha$)</th>
<th>Posterior Mean</th>
<th>Posterior SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TOPIC_ENTROPY$</td>
<td>0.064***</td>
<td>(0.021)</td>
</tr>
<tr>
<td>$NUM_WORDS$</td>
<td>0.019</td>
<td>(0.014)</td>
</tr>
<tr>
<td>$BRAND$</td>
<td>0.064**</td>
<td>(0.025)</td>
</tr>
<tr>
<td>$LOCATION$</td>
<td>-0.041*</td>
<td>(0.025)</td>
</tr>
<tr>
<td>$LOG_TRANS$</td>
<td>-0.012</td>
<td>(0.010)</td>
</tr>
<tr>
<td>$LOG_IMP$</td>
<td>-0.019**</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$AVG_NUM_ADS$</td>
<td>0.372***</td>
<td>(0.009)</td>
</tr>
<tr>
<td>$AVG_AD_QUALITY$</td>
<td>28.722***</td>
<td>(0.412)</td>
</tr>
</tbody>
</table>

$\mu_0^{(\alpha)} = -2.654***$ (0.094)

***, **, and * indicate a 99%, 95%, and 90% significance level. The same notation is followed for other tables in the paper.
Second, we observe that the presence of a brand name in the keyword significantly increases the likelihood of a click. This finding is consistent with the results shown by Rutz and Bucklin (2011), who use data from a single advertiser to show that the presence of a brand name increases the CTR. The negative coefficient for LOCATION indicates that keywords that have location information are less likely to get clicked than keywords that do not contain such information. Analysis on several search engines reveals that keywords that contain location information tend to result in organic search results that are relevant to the user’s search intent. As a result, search ads provide little additional information and hence are less likely to be clicked. However, the coefficient of LOG_TRANS is not significant, indicating that transactional intent does not seem to affect the baseline CTR.

Third, consistent with Jerath et al. (2013) and Rutz and Trusov (2011), we observe that as the number of impressions for a keyword increase, the CTR decreases. Jerath et al. (2013) suggest that more popular keywords (with higher impressions) are used by consumers early on in their search process, and due to the low level of involvement, have fewer clicks. By contrast, less popular keywords are used by consumers in later stages of their purchase deliberation, and are more likely to generate clicks. Regarding how the ad characteristics affect the keyword performance, the average number of ads displayed for a keyword has a significantly positive impact on the CTR. One might argue this result is driven by endogeneity – keywords that are more likely to be clicked attract more advertisers. To address this concern, we analyze the impression-level data for a few keywords that have considerable variation in the number of ads across impressions, and obtain qualitatively similar results. The quality of ads has a positive effect on the CTR, which seems fairly intuitive.

Finally, the estimate for $\mu_0^{(a)} (-2.645)$ indicates that on average the baseline CTR is fairly low. Note that because we have mean centered all the covariates, $\mu_0^{(a)}$ can be interpreted as the average $\bar{\alpha}_{kt}$ (see subsection 5.2) for a keyword with covariates at mean values. The topic-specific intercept terms for the 20 categories are presented in Figure 6. Considerable variation occurs in the keyword performance across categories as indicated by the intercept terms. Overall, keywords in categories such as “baby” (e.g., “babies r us”), “listing” (e.g., “house for sale”), “finance” (e.g., “unsecured loans”), and “travel” (e.g., “cheap vacations”) are likely to receive more clicks than keywords in categories such as “TV” (e.g., “abc tv shows”) and “sport” (e.g., “redsox”). This finding indicates that consumers exhibit different search behaviors across different product categories. Several reasons may drive the difference across categories. Consumers conducting television- or sports-related search might search less because their informational or transactional needs are met by organic search results, or the ads are not as relevant.

---

8 A rough calculation of the baseline CTR using logistic function: $\frac{e^{(-2.645)}}{1 + e^{(-2.645)}} = 0.066$. 

On the other hand, consumers using baby- or listing-related keywords might be willing to search more, or organic search listings do not meet their search needs. Because we do not have the organic search results or further characterization of the categories used in our analysis, providing an explanation for the difference in keyword performance across the categories is difficult, and we leave this avenue as a direction for future research. Advertisers can use the result shown here to decide the portfolio of keywords for search advertising. Keywords that belong to categories such as “baby” and “listing” might lead to more clicks, whereas keywords that belong to “TV” or “sport” might generate fewer clicks. An advertiser can use these insights to appropriately choose his advertising strategy.

6.2.2 Effect on $\gamma$

We present the effect of different factors on the decay in CTR with position in Table 6.2.2. First, we observe that topic entropy has a negative and significant impact on the decay parameter $\gamma$. That is, keywords that have higher topic entropy seem to have a lower $\gamma$ and witness a larger decrease in CTR with position. This finding indicates that although on average, consumers are more likely to click ads associated with more ambiguous keywords, especially from the top positions, they are less likely to click ads that are positioned lower on the screen once they start the search. In other words, during the
sequential search process, consumers are more likely to lose their search interests and stop exploring the choice set further under more ambiguous keywords. Reduced consumer search activity associated with ambiguous ads can be attributed to search costs theory (e.g., Stigler 1961) and information overload theory (e.g., Iyengar and Lepper 2000, Dhar and Simonson 2003, and Kuksov and Villas-Boas 2010). The well-established search costs literature has demonstrated that higher search costs can lead to lower search intensity (e.g., Stigler 1961; Weitzman 1979). Meanwhile, previous research has shown both theoretical and empirical evidence that information overload can discourage consumers and lead to consumer search termination due to loss of interests in the search results (e.g., Iyengar and Lepper 2000, Kuksov and Villas-Boas 2010, and Ghose et al. 2013). In the context of sponsored search, once consumers start clicking ads from top to bottom, they start to learn the diverse quality of the ads in sequence. Moreover, they also learn their search costs of searching each of these ads. Due to the high heterogeneity in the quality of the ads generated by the ambiguous keywords, consumers might find it cognitively costly to process these ads to pick the relevant ones (hence high search costs). In addition, a highly diverse set of ads might introduce an increasing amount of new information to consumers during the search process. Such information may lead to an overload for consumers due to their cognitive limitation and non-negligible search costs. Therefore, when keywords are more ambiguous, we notice a significant reduction in the probability of clicking on the ads as consumers proceed down the list. This finding is also highly consistent with the effects of knowledge uncertainty found by Urbany et al. (1989). In particular, keywords with high topic entropy, and hence high ambiguity, may occur in search contexts where consumers’ knowledge uncertainty is high. The lack of knowledge can in turn increase the barrier for consumers to evaluate choices, leading to much higher search costs for consumers during the search process. This interpretation is in line with the search costs theory where higher search costs can lead to lower search intensity.

Second, we find that other keyword characteristics can also affect the decay of CTR with screen position. In particular, keywords that are longer and transaction-related are associated with a higher value of $\gamma$, meaning the corresponding CTR tends to decrease at a slower rate given a lower screen position. In other words, consumers are more likely to traverse further down the list in the context of specific or transaction-related keywords. The intuition behind this result can be explained as follows: consumers who search with a more specific keyword or transactional intent are primed toward transactional actions, and thus may be willing to click on more ads to find the best alternative by engaging in activities such as price or feature comparison. By contrast, we find that position effect tends to be much stronger for brand- or location-specific keywords. This finding suggests that when consumers search for
Table 6: Estimation Results for $\gamma$

<table>
<thead>
<tr>
<th></th>
<th>Decay Parameter ($\gamma$)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posterior Mean</td>
<td>Posterior SD</td>
<td></td>
</tr>
<tr>
<td>TOPIC_ENTROPY</td>
<td>-0.145*** (0.023)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUM_WORDS</td>
<td>0.055*** (0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAND</td>
<td>-0.161*** (0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOCATION</td>
<td>-0.193*** (0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>0.047*** (0.011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOG_IMP</td>
<td>-0.059*** (0.011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG_NUM_AD</td>
<td>0.256*** (0.010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG_AD_QUALITY</td>
<td>-7.170*** (0.540)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_0^{(\gamma)}$</td>
<td>-0.263* (0.144)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

brand- or location-related information, they might not explore ads that are comparatively lower in the list. This faster decay with position associated with brand-related keywords might be driven by the fact that firms bidding on their own keywords associated with their own brand names try to appear at the top positions. As a result, consumers tend not to click on ads that are below the “branded” ad, which leads to a smaller $\gamma$. We also observe that as the average number of ads increases, consumers tend to click on more ads. Interestingly, consumers search less as the ad quality associated with a keyword increases, which might be a consequence of the first few ads being sufficient to satisfy their search or transactional intent.

The estimate of $\mu_0^{(\gamma)}$ suggests the average decay of CTR with position is significant. To understand the heterogeneous impact of position on CTR across topics, we demonstrate the topic-specific intercepts for $\gamma$ across 20 different categories in Figure 7. We find the corresponding intercepts for $\gamma$ are highest for topics such as “adult,” “home” (e.g., “bathroom vanity”), and “style” (e.g., “wedding dresses”), suggesting that consumers who are interested in those topics may be more likely to click on ads at lower screen positions. By contrast, for topics related to “baby,” “finance,” and “travel,” CTR decreases at a faster rate with position.
6.2.3 Covariance Matrices $\Phi$ and $\Omega_0$

Table 7 shows the estimates for the covariance matrix between $\tilde{\alpha}_{kt}$ and $\tilde{\gamma}_{kt}$, which capture unobserved keyword-position heterogeneity (equation 4). The estimated variances are significant, suggesting considerable variation across keywords. The covariance estimate between $\tilde{\alpha}_{kt}$ and $\tilde{\gamma}_{kt}$ is significant and negative, indicating a negative relationship between $\alpha$ (baseline CTR) and $\gamma$ (decay parameter).

Similarly, the estimates for the covariance matrix between $\beta^{(\alpha)}_{0t}$ and $\beta^{(\gamma)}_{0t}$, which capture topic-specific heterogeneity (Equation 6), are reported in Table 8. The estimated variances for $\beta^{(\alpha)}_{0t}$ and $\beta^{(\gamma)}_{0t}$ are also significant, suggesting that both the baseline CTR and the decay parameter vary across topics. The estimated covariance between $\beta^{(\alpha)}_{0t}$ and $\beta^{(\gamma)}_{0t}$ is negative and significant, which can also be observed by comparing Figure 6 and Figure 7. For example, topics such as “baby” and “finance” tend to attract more clicks at the top position, but the CTR tends to drop quickly with position, whereas topics such as “TV” and “home” tend to attract fewer clicks at the top position, but the CTR decays at a slower rate.
Table 7: Estimation Results for $\Phi$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\alpha}_{kt}$</th>
<th>$\hat{\gamma}_{kt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\alpha}_{kt}$</td>
<td>0.173*** (0.066)</td>
<td>-0.155*** (0.078)</td>
</tr>
<tr>
<td>$\hat{\gamma}_{kt}$</td>
<td>-0.155*** (0.078)</td>
<td>0.408*** (0.147)</td>
</tr>
</tbody>
</table>

Table 8: Estimation Results for $\Omega_0$

<table>
<thead>
<tr>
<th></th>
<th>$\beta_{\alpha}^{(\alpha)}$</th>
<th>$\beta_{\alpha}^{(\gamma)}$</th>
<th>$\beta_{\gamma}^{(\alpha)}$</th>
<th>$\beta_{\gamma}^{(\gamma)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{\alpha}^{(\alpha)}$</td>
<td>0.390*** (0.011)</td>
<td>-0.249*** (0.011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{\gamma}^{(\gamma)}$</td>
<td>-0.249*** (0.011)</td>
<td>0.450*** (0.014)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 6.3 Summary of Main Results

The analysis presented here suggests that consumers follow a two-step process while engaging in sponsored search. Their behavior while deciding whether to click on search ads differs considerably from their behavior once they have started search, as evidenced by the differing effect of keyword characteristics on $\alpha$ and $\gamma$. For example, keyword entropy has two opposing effects on the performance of ads associated with a keyword. Our finding suggests that keyword contextual ambiguity can lead to a higher baseline CTR; however, it also implies the CTR decreases more sharply with position. This dual effect of entropy on keyword performance is closely related to prior research by Urbany et al. (1989) that divides the uncertainty consumers face into knowledge uncertainty and choice uncertainty. Choice uncertainty captures a consumer’s uncertainty in comparing alternatives and leads to more search, whereas knowledge uncertainty reflects customer uncertainty about the dimensions of comparison and leads to less search. In our case, higher entropy leads to higher choice uncertainty overall, hence increasing overall consumer search. However, once consumers start a search, higher entropy also leads to higher knowledge uncertainty from each position as the consumers become confused by incoherent ads, resulting in deceased search activity and thus a faster decay in CTR with position. The overall effect of keyword entropy on CTR for ads at various positions is a combination of these opposing effects. In our analysis, we observe that the decrease in the decay with position always dominates the increase in the baseline CTR. For example, the ads at positions 2-8 for a high entropy keyword tend to get a smaller number of clicks as compared to ads for a keyword with lower entropy. This finding indicates that firms should either advertise on less ambiguous keywords or bid aggressively on the top position for ambiguous keywords if they are interested in increasing traffic to their websites. Figure 8 illustrates
how the decay of CTR varies by entropy. We use the CTR at each position of a sample keyword with mean entropy as the baseline (i.e., normalized to zero), and plot the CTR for high- and low-entropy keywords at each position relative to the mean-entropy keyword. As seen in Figure 8, relative to the mean-entropy keyword, the low-entropy keyword starts with lower CTR at the top position, but has higher CTR at positions 2-8. By contrast, the high-entropy keyword has higher CTR at the top position than the mean-entropy keyword, but lower CTR at lower positions. This two-step process of consumer search is also exemplified by the different effects of BRAND and NUM_WORDS on $\alpha$ and $\gamma$.

Figure 9 illustrates how CTR changes with position by topic. We use different colors to demonstrate different levels of CTR. Darker red colors represent higher CTR values and light yellow colors represent lower CTR values. Our results suggest the position effect on CTR is heterogeneous across different topics. For example, baby-related keywords tend to attract higher CTR than listing-related keywords at the top position, but the CTR decreases quickly at lower positions. This pattern we observe for baby-related keywords might be driven by the fact that consumers are extremely interested in baby-related ads, which leads to a high CTR at the first position, but their search intent is satisfied after the first few clicks. On the other hand, searches related to “home,” “style,” “clothing,” and “adult” start out with fewer clicks on ads in the top position but continue to receive clicks on ads in lower positions. In fact, we observe that the search depth for “adult” is the largest among all the categories considered in our
The large-scale, cross-category analysis presented here enables us to generate insights that depart from well-established results in the sponsored search literature. First, we can leverage data from multiple advertisers to generate insights that cannot be derived using data from a single advertiser. For example, we observe that ads associated with branded keywords tend to have smaller CTR at lower positions than generic keywords. At first, this finding might seem contrary to the results shown by Rutz and Bucklin (2011), who use data from a single advertiser to show that the presence of a brand name always increases CTR. To illustrate why data from a single advertiser might lead to such a result, consider the following example. An advertiser (Hilton) bids on two keywords, “hotel NYC” and “Hilton NYC,” and shows an identical ad for both the keywords. A consumer who uses the keyword “Hilton NYC” is looking for a Hilton property and is more likely to click on an ad by Hilton. On the other hand, a consumer searching for “hotel NYC” is looking for hotels in New York City and might not have a strong preference for Hilton, reducing the likelihood of clicking on an ad by Hilton. Ceterius paribus, the advertiser will see more clicks for the keyword “Hilton NYC” than “hotel NYC,” and analyzing data from only one advertiser might lead to the conclusion that branded keywords always lead to more clicks. Our analysis,
on the other hand, is based on a large sample of keywords across different advertisers, and we believe our findings complement the earlier findings. We conclude that, in general, keywords that contain brand names perform better than keywords that do not; however, a sharp decrease occurs in the CTR (with position) for branded keywords. This insight is manifested in the positive effect of BRAND on $\alpha$ but a negative effect on $\gamma$, resulting in an overall lower CTR for branded keywords. Second, data from all the advertisers for a keyword help us estimate the effect of keyword characteristics on the baseline CTR and the decay parameter separately. Such a model allows us to measure the differential effect of factors such as LOG_TRANS and the topic membership on $\alpha$ and $\gamma$, and gain a better insight of customer search behavior in the context of sponsored search.

7 Robustness Checks and Model Comparison

7.1 Alternative Model Specifications

To check the prediction performance of our proposed model, we compare our model against several other alternative models. For convenience, we refer to Model 1 as our proposed model. Model 2 has the same model specification as Model 1, but it excludes topic entropy in equation 5. Model 3 is an alternative model without topic entropy or topic-specific intercepts. The main difference between Model 2 and Model 3 is that Model 3 does not incorporate topic heterogeneity. Therefore, in Model 3, we drop the topic subscript in equation 4 as $(\beta_0^{(\alpha)}, \beta_0^{(\gamma)})^\prime \sim MVN(\mu_0, \Omega_0)$.

Model 4 is a logit model with topic entropy and topic-specific intercepts, where the interaction effects between position and covariates are linearly additive instead of nonlinear. Specifically, in Model 4, we assume $CTR_{kpt} = \frac{\exp(\tilde{u}_{kpt})}{1+\exp(\tilde{u}_{kpt})}$, where $\tilde{u}_{kpt} \sim N(\mu_{kpt}, \sigma^2)$. We further assume $\mu_{kpt} = \beta_0t + X_k^\prime \beta_1 + (pX)_k^\prime \beta_2 + \beta_3p$, where $X_k^\prime$ is a vector of characteristics for keyword $k$ and $p$ denotes position.

To compare alternative model specifications, we use 70% of the keywords as the estimation sample, and the remaining 30% of the keywords as the holdout sample (see Appendix C for summary statistics for the two samples). We estimate the four alternative specifications using MCMC with the estimation sample and compare the deviance information criteria (DIC; Spiegelhalter et al. 2002). Table 9 summarizes the differences among the four models and reports model fit statistics. As can be seen, Model 1 provides the smallest DIC among the four alternative models.
Table 9: Summary of Alternative Models

<table>
<thead>
<tr>
<th>Description</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full model</td>
<td>No entropy</td>
<td>No entropy or topic heterogeneity</td>
<td>Logit model</td>
</tr>
<tr>
<td><strong>Topic-specific Intercept</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Included</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Entropy Included</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Control Variables</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Included</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DIC</strong></td>
<td>222,822</td>
<td>354,369</td>
<td>227,568</td>
<td>330,816</td>
</tr>
</tbody>
</table>

7.2 Model Comparison: Prediction Performance

To compare the prediction performance of our model (Model 1) against the alternative models, we perform both in-sample and out-of-sample prediction. Based on the estimation results using the estimation sample, we predict CTR and the number of clicks for each keyword at each position, and use root-mean-square error (RMSE) to evaluate prediction accuracy. Table 10 reports the RMSE for each sample for each model based on the CTR prediction, and Panel B reports both the RMSE based on number of clicks. In all cases, our model (Model 1) outperforms the other three models, suggesting that incorporating topic heterogeneity, topic entropy, and non-linear position effects provide better prediction accuracy.

Table 10: Model Comparison: Prediction Performance based on RMSE

<table>
<thead>
<tr>
<th></th>
<th>A: Model Fit Based on CTR</th>
<th>B: Model Fit Based on Number of Clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>in-sample</td>
<td>out-of-sample</td>
</tr>
<tr>
<td>Model 1</td>
<td>0.0328</td>
<td>0.0324</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.0329</td>
<td>0.0324</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.0334</td>
<td>0.0331</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.0346</td>
<td>0.0341</td>
</tr>
</tbody>
</table>
8 Discussion and Conclusions

In this paper, we study the effect of semantic characteristics on keyword click-through rate and try to provide insights into consumer behavior in the context of search advertising. We analyze data from a major US search engine for 12,790 distinct keywords spanning 4.6 million impressions. Our paper is unique in several ways. First, to our knowledge, this dataset is the most extensive one used in the sponsored search literature that includes individual-level data across multiple product categories and advertisers. Most of the prior research in this area has used data from a single advertiser or a few keywords from search engines. Second, we are able to exploit machine learning techniques to generate semantic characteristics of keywords given the large volume of keywords available to us. The large number of keywords helps us create meaningful categorization without any manual intervention. Using text mining techniques, we are able to generate keyword characteristics such as presence of brand name and location, length of keywords, and whether a keyword is transactional, and study the impacts of these factors on consumer click behavior. Third, we introduce a new keyword characteristic – topic entropy – that measures the ambiguity in the semantic meaning of a keyword. We find that topic entropy has a significant impact on the CTR. Our results suggest topic entropy and keyword categories are significant predictors of keyword performance and potentially affect consumer click behavior.

Our results indicate that when consumers search using ambiguous keywords, they are more likely to click on sponsored ads in top positions. However, consumers who click on ads associated with these ads tend to search less aggressively and click fewer ads in lower positions. More specifically, CTR decays faster with position for ambiguous keywords as compared to precise keywords. We also observe that consumer behavior varies significantly across categories. Categories such as TV and sports receive fewer clicks, whereas categories such as style and travel receive significantly more clicks. We extend the prior understanding of brand- and location-specific keywords. Our results suggest consumers are “in general” less likely to click on ads when searching for location related information, and are less likely to conduct intensive search when searching for brand information. These counterintuitive results (as compared to previous literature) that shed further light on consumer behavior are derived because we are able to capitalize on data across different advertisers, whereas most prior studies use data from a single advertiser. We also observe that as the search volume of the keywords increases, the number of clicks decreases. In addition, as the number of ads for an impression increases, consumers are likely to click on more ads.

Our research has several managerial implications for advertisers and search engines. First, advertisers
who are interested in driving traffic through clicks should be cautious about adding ambiguous keywords to their portfolios. Furthermore, if they need to choose some ambiguous keywords, they might be better off bidding aggressively and trying to attain higher positions. The search engines should also be cognizant of the topic ambiguity while designing their sponsored search strategies. For example, search engines may incorporate keyword characteristics while evaluating ad quality. Second, given that consumers behave differently across categories, advertisers can design their advertising strategies by allocating resources appropriately across products. Third, advertisers can use insights about how consumers respond to different keyword characteristics in designing better campaigns.

This paper incorporates new machine learning techniques in the literature that are instrumental in deriving new insights. We hope this research can pave the way for interesting future research that addresses the limitation of our current work. One major limitation of our analysis is the lack of ad information. For example, we don’t know ad characteristics such as ad copy, bid, and landing page. Although we try to compute some ad measures from our data and use these measures as proxies to control for ad-level differences in our analysis, providing richer insights with ad-level data is possible. Another limitation of this paper is the lack of post-click or conversion activity in the data. Therefore, our measurement of keyword performance is limited only to CTR. From the perspective of a search engine, CTR is a more relevant metric, because advertisers only pay when a click-through occurs. However, advertisers are not only interested in CTR, but they are also interested in conversion rates as well as other non-transactional benefits such as increased awareness. A richer dataset can be used to address this issue. We believe that sponsored search advertising continues to evolve as an interesting area of marketing research and this paper can contribute both methodologically and theoretically to this growing literature.

References


V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet


S. Yang and A. Ghose. Analyzing the relationship between organic and sponsored search advertising:

### Appendix A. Summary of Empirical Studies on Sponsored Search Advertising

<table>
<thead>
<tr>
<th>Paper</th>
<th>Goal</th>
<th>Data Source</th>
<th>Industry</th>
<th>Level of detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarwal et al. (2011)</td>
<td>Impact of position on click-through, conversion and profit</td>
<td>Advertiser</td>
<td>Retail</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Animesh et al. (2011)</td>
<td>Impacts of a firm’s positioning strategy, ad position, and competition on ad performance</td>
<td>Advertiser</td>
<td>Mortgage</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Chan et al. (2011)</td>
<td>Measuring the value of customers acquired from sponsored search</td>
<td>Advertiser</td>
<td>Lab supplies</td>
<td>Individual</td>
</tr>
<tr>
<td>Ghose and Yang (2009)</td>
<td>Impact of keyword attributes on click-through and conversion</td>
<td>Advertiser</td>
<td>Retail</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Goldfarb and Tucker (2011)</td>
<td>Online and offline advertising channel substitution</td>
<td>Advertiser</td>
<td>Legal service</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Jerath et al. (2013)</td>
<td>Impact of keyword popularity on click performance</td>
<td>Search engine</td>
<td></td>
<td>Individual</td>
</tr>
<tr>
<td>Jezierski and Segal (2012)</td>
<td>Quantifying rational user experience and externalities among ads</td>
<td>Search engine</td>
<td></td>
<td>Individual</td>
</tr>
<tr>
<td>Rutz and Bucklin (2011)</td>
<td>Spill-over from generic to branded keywords</td>
<td>Advertiser</td>
<td>Hotel</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Authors</td>
<td>Title</td>
<td>Advertiser</td>
<td>Ad Format</td>
<td>Aggregate</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>------------</td>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Rutz and Trusov (2011)</td>
<td>Effects of ad attributes on ad performance</td>
<td>Advertiser</td>
<td>Ringtone</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Rutz et al. (2011)</td>
<td>Quantifying indirect effects of paid search</td>
<td>Advertiser</td>
<td>Automotive</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Rutz et al. (2012)</td>
<td>Impact of ad position on conversion performance</td>
<td>Advertiser</td>
<td>Hotel</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Yang and Ghose (2010)</td>
<td>Relationship between organic and sponsored search</td>
<td>Advertiser</td>
<td>Retail</td>
<td>Aggregate</td>
</tr>
<tr>
<td>Yao and Mela (2011)</td>
<td>Modeling user, advertiser, and search engine interaction</td>
<td>Search Engine</td>
<td>Music management</td>
<td>Individual</td>
</tr>
</tbody>
</table>
## Appendix B. Correlation among Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>IMP</th>
<th>CLICKS</th>
<th>CLICK_DEPTH</th>
<th>CTR</th>
<th>TOPIC_ENTROPY</th>
<th>NUM_WORDS</th>
<th>BRAND</th>
<th>LOCATION</th>
<th>LOG_TRANS</th>
<th>AVG_AD_QUALITY</th>
<th>AVG_NUM_ADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLICKS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLICK_DEPTH</td>
<td>-0.029</td>
<td>0.209</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTR</td>
<td>-0.006</td>
<td>0.298</td>
<td>0.859</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOPIC_ENTROPY</td>
<td>-0.001</td>
<td>-0.011</td>
<td>-0.114</td>
<td>-0.070</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUM_WORDS</td>
<td>-0.031</td>
<td>-0.040</td>
<td>0.020</td>
<td>0.005</td>
<td>-0.361</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAND</td>
<td>0.062</td>
<td>0.108</td>
<td>0.043</td>
<td>0.065</td>
<td>-0.031</td>
<td>0.066</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOCATION</td>
<td>-0.012</td>
<td>-0.016</td>
<td>-0.086</td>
<td>-0.070</td>
<td>-0.017</td>
<td>0.161</td>
<td>0.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>-0.002</td>
<td>0.100</td>
<td>0.304</td>
<td>0.265</td>
<td>0.017</td>
<td>0.011</td>
<td>0.171</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG_AD_QUALITY</td>
<td>-0.007</td>
<td>0.293</td>
<td>0.760</td>
<td>0.869</td>
<td>-0.081</td>
<td>0.010</td>
<td>0.084</td>
<td>-0.064</td>
<td>0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG_NUM_ADS</td>
<td>-0.112</td>
<td>0.012</td>
<td>0.525</td>
<td>0.242</td>
<td>-0.107</td>
<td>-0.010</td>
<td>0.006</td>
<td>-0.017</td>
<td>0.218</td>
<td>0.245</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C. Summary Statistics: Estimation Sample and Holdout Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>A: Estimation Sample</th>
<th>B: Holdout Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std Dev</td>
</tr>
<tr>
<td>IMP</td>
<td>347.615</td>
<td>3443.453</td>
</tr>
<tr>
<td>CLICKS</td>
<td>49.972</td>
<td>197.152</td>
</tr>
<tr>
<td>CLICKS_DEPTH</td>
<td>0.186</td>
<td>0.167</td>
</tr>
<tr>
<td>AVG_CTR</td>
<td>0.027</td>
<td>0.025</td>
</tr>
<tr>
<td>TOPIC_ENTROPY</td>
<td>1.604</td>
<td>0.450</td>
</tr>
<tr>
<td>NUM_WORDS</td>
<td>1.740</td>
<td>0.679</td>
</tr>
<tr>
<td>BRAND</td>
<td>0.167</td>
<td>0.373</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.131</td>
<td>0.337</td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>2.174</td>
<td>1.074</td>
</tr>
<tr>
<td>AVG_NUM_ADS</td>
<td>6.606</td>
<td>1.210</td>
</tr>
<tr>
<td>AVG_AD_QUALITY</td>
<td>0.026</td>
<td>0.023</td>
</tr>
</tbody>
</table>
Appendix D. The Gibbs Sampling Procedure

We estimate our hierarchical Bayesian model using a Gibbs sampling procedure, which samples parameters iteratively from their conditional distribution given the data and all other parameters. Note that we model a keyword $k$’s CTR conditional on the membership in a topic $t$, with membership probabilities $\hat{\theta}_{kt}$ estimated from a topic model. Instead of analytical derivation of the unconditional CTR, we use a data augmentation approach by simulating topic assignment.

First, the posterior distribution of all parameters can be written as follows. (For convenience, we will not explicitly mention data in conditional distributions.)

$$f(\{\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}\}, \beta^{(a)}_t, \beta^{(\gamma)}_t, \mu_0, \Phi, \Omega_0, \{t\}|data, \{\hat{\theta}_k\}; \Sigma^{(a)}, \Sigma^{(\gamma)}, V^\Phi, \nu^\Phi, V^\Omega, \nu^\Omega) \propto f(\{CLICKS_{kp}, IMP_{kp}\}|\{\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}\}, \{topic_k = t\}) \cdot f(\{topic_k = t\}|\{\hat{\theta}_k\})$$

$$\cdot f(\{\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}\}|\beta^{(a)}_t, \beta^{(\gamma)}_t, \beta^{(a)}_t, \beta^{(\gamma)}_t, \Phi) \cdot f(\{\beta^{(a)}_t, \beta^{(\gamma)}_t\}|\mu_0, \Theta_0)$$

$$\cdot f(\beta^{(a)}_t|\Sigma^{(a)}) \cdot f(\beta^{(\gamma)}_t|\Sigma^{(\gamma)}) \cdot f(\mu_0|\Theta_0) \cdot f(\Phi|\nu^\Phi, \nu^\Omega) \cdot f(\Theta_0|\nu^\Omega, \nu^\Omega),$$

where $\{\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}\}, \beta^{(a)}_t, \beta^{(\gamma)}_t, \beta^{(a)}_t, \beta^{(\gamma)}_t, \mu_0, \Phi,$ and $\Omega_0$ are parameters of interests. $\{t\}$ is topic assignment to be sampled from a multinomial distribution with probabilities $\hat{\theta}_k$ for keyword $k$. $\Sigma^{(a)}, \Sigma^{(\gamma)}, V^\Phi, \nu^\Phi, V^\Omega,$ and $\nu^\Omega$ are parameters specified for the prior distributions.

Next, we describe the Gibbs sampling procedure.

**Step 1**

Draw $\text{topic}_k = t | \hat{\theta}_k \sim \text{Multinomial}(\hat{\theta}_k)$ for each keyword $k$.

**Step 2**

Draw $\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}|\beta^{(a)}_t, \beta^{(\gamma)}_t, \beta^{(a)}_t, \beta^{(\gamma)}_t, \Phi$ for each keyword $k$ in topic $t$.

$$f(\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}|\beta^{(a)}_t, \beta^{(\gamma)}_t, \beta^{(a)}_t, \beta^{(\gamma)}_t, \Phi) \propto \left(\frac{\exp(\tilde{\alpha}_{kt})}{1 + \exp(\tilde{\alpha}_{kt})}\right)^{p-1}CLICKS_{kp}(1 - \frac{\exp(\tilde{\alpha}_{kt})}{1 + \exp(\tilde{\alpha}_{kt})})^{p-1}IMP_{kp} - CLICKS_{kp} \cdot f(\tilde{\alpha}_{kt}, \tilde{\gamma}_{kt}|\mu_{kt}, \Phi),$$

where $\mu_{kt} = (\mu^{(a)}_{kt}, \mu^{(\gamma)}_{kt}) = (\beta^{(a)}_t + X_k\beta^{(a)}_t, \beta^{(\gamma)}_t + X_k\beta^{(\gamma)}_t)$.

**Step 3**
Draw $\beta_{0t}^{(a)}, \beta_{0t}^{(\gamma)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(a)}, \beta^{(\gamma)}, \mu_0, \Omega_0$ for each topic $t$.

\[ f(\beta_{0t}^{(a)}, \beta_{0t}^{(\gamma)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(a)}, \beta^{(\gamma)}, \mu_0, \Omega_0) \propto f(\{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\} \mid \mu_k) \cdot f(\{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\} \mid \mu_0, \Omega_0), \]

where $\mu_{kt} = (\mu_{kt}^{(a)}, \mu_{kt}^{(\gamma)})' = (\beta_{0t}^{(a)} + X_k'\beta^{(a)}, \beta_{0t}^{(\gamma)} + X_k'\beta^{(\gamma)})'$.

**Step 4**

Draw $\beta^{(a)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(\gamma)}, \Sigma^{(a)}$.

\[ f(\beta^{(a)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(\gamma)}, \Sigma^{(a)}) \propto f(\{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\} \mid \mu_k) \cdot f(\beta^{(a)} \mid \Sigma^{(a)}), \]

where $\mu_{kt} = (\mu_{kt}^{(a)}', \mu_{kt}^{(\gamma)}') = (\beta_{0t}^{(a)} + X_k'\beta^{(a)}, \beta_{0t}^{(\gamma)} + X_k'\beta^{(\gamma)})'$.

**Step 5**

Draw $\beta^{(\gamma)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(a)}, \Sigma^{(\gamma)}$.

\[ f(\beta^{(\gamma)} \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \beta^{(a)}, \Sigma^{(\gamma)}) \propto f(\{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\} \mid \mu_k) \cdot f(\beta^{(\gamma)} \mid \Sigma^{(\gamma)}), \]

where $\mu_{kt} = (\mu_{kt}^{(a)}', \mu_{kt}^{(\gamma)}') = (\beta_{0t}^{(a)} + X_k'\beta^{(a)}, \beta_{0t}^{(\gamma)} + X_k'\beta^{(\gamma)})'$.

**Step 6**

Draw $\mu_0 \mid \{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\}, \Sigma_0$.

\[ f(\mu_0 \mid \{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\}, \Sigma_0) \propto f(\{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\} \mid \mu_0, \Omega_0) \cdot f(\mu_0 \mid \Sigma_0). \]

**Step 7**

Draw $\Phi \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\}, \beta^{(a)}, \beta^{(\gamma)}$.

\[ f(\Phi \mid \{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\}, \{\beta_{0t}^{(a)}\}, \{\beta_{0t}^{(\gamma)}\}, \beta^{(a)}, \beta^{(\gamma)}) \propto f(\{\tilde{\alpha}_{kt}\}, \{\tilde{\gamma}_{kt}\} \mid \mu_k) \cdot f(\Phi \mid V^\Phi, \nu^\Phi), \]
where \( \mu_{kt} = (\mu_{kt}^{(\alpha)}, \mu_{kt}^{(\gamma)})' = (\beta_{0t}^{(z)} + X'k\beta^{(z)}, \beta_{0t}^{(w)} + X'k\beta^{(w)})' \).

**Step 8**

Draw \( \Omega_0|\{\beta_{0t}^{(z)}\}, \{\beta_{0t}^{(w)}\}, \mu_0, V, \nu \).

\[
f(\Omega_0|\{\beta_{0t}^{(z)}\}, \{\beta_{0t}^{(w)}\}, \mu_0, V, \nu) \propto f(\{\beta_{0t}^{(z)}\}, \{\beta_{0t}^{(w)}\}|\mu_0, \Omega_0) \cdot f(\Omega_0|V, \nu).
\]