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

Jing Gong[†], Vibhanshu Abhishek[‡], Beibei Li[‡]
[†]Temple University, [‡]Carnegie Mellon University

August 11, 2017

Abstract

In this paper, we explore how keyword ambiguity can affect search advertising performance. Consumers arrive at search engines with diverse interests, which are often unobserved and nontrivial to predict. The search interests of different consumers may vary even when they are searching using the same keyword. In our study, we propose an automatic way of examining keyword ambiguity based on probabilistic topic models from machine learning and computational linguistics. We examine the effect of keyword ambiguity on keyword performance using a hierarchical Bayesian approach that allows for topic-specific effects and nonlinear position effects, and jointly models click-through rate (CTR) and ad position (rank). We validate our study using a novel data set from a major search engine that contains information on consumer click activities for 2,625 distinct keywords across multiple product categories from 10,000 impressions. We find that consumer click behavior varies significantly across keywords, and such variation can be partially explained by keyword ambiguity. Specifically, higher keyword ambiguity is associated with higher CTR on top-positioned ads, but also a faster decay in CTR with screen position. Therefore, the overall effect of keyword ambiguity on CTR varies across positions. Our study provides implications for advertisers to improve the prediction of keyword performance by taking into account keyword ambiguity and other semantic characteristics of keywords. It can also help search engines design keyword planning tools to aid advertisers when choosing potential keywords.

Keywords: *Sponsored search advertising, topic models, keyword ambiguity, machine learning, hierarchical Bayesian model.*

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

INTRODUCTION

With the growing pervasiveness of consumer search for relevant information and products via search engines, search advertising has become an important marketing channel for businesses.[1] In 2016, search advertising generated revenues of $35.0 billion and accounted for 48% of online advertising revenues (Interactive Advertising Bureau 2017). Most forms of online advertising offer a more effective way of targeting customers as compared to traditional advertising. However, search advertising considerably outperforms other forms of online advertising such as display or social media advertising on metrics such as return on investment, click-through rate (CTR), and conversion rate (Szymanski and Lee 2006). 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 (Agarwal et al. 2011).

When a consumer issues a query on a search engine using a keyword, for example, "hotels nyc," the search engine identifies and returns two lists of search results: a list of "organic" search results recommending web pages relevant to the keyword "hotels nyc," and a list of "sponsored" ads by the advertisers who are bidding on the keyword "hotels nyc." The sponsored ads are determined 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 interests (as indicated by the keywords) considerably increases the likelihood that they will click on these ads (Agarwal et al. 2011). 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 consumers when they are engaged in unrelated activities, and the consumers might not have an

[1] Search advertising is also known as "paid search" or "sponsored search advertising."

immediate search/purchase intent when they encounter these ads. The pull-based nature of sponsored search ads leads to their relatively high performance as compared to other forms of advertising.

Even though a keyword provides an indication of a consumer's search interest, consumers with varied interests might use the same keyword for searching. For example, a consumer who searches for the keyword "Mars" may be interested in astronomy and the planet Mars, 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, checking about a Las Vegas hotel, or interested in a 1977 movie. Although search engines are trying to predict consumers' interests, a particular consumer's search interest 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 that the search engine faces in predicting consumers' search interests.

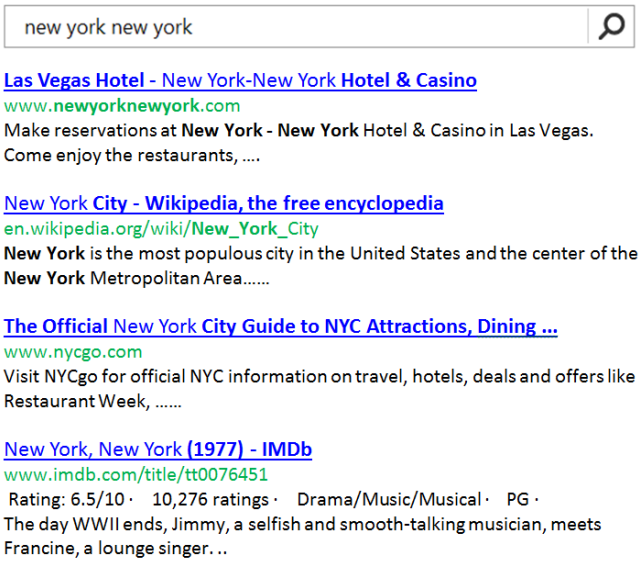

Figure 1: Organic Search Results of the Keyword "new york, new york" on a Major Search Engine.

Since the same keyword can refer to different search interests, competing advertisers might have

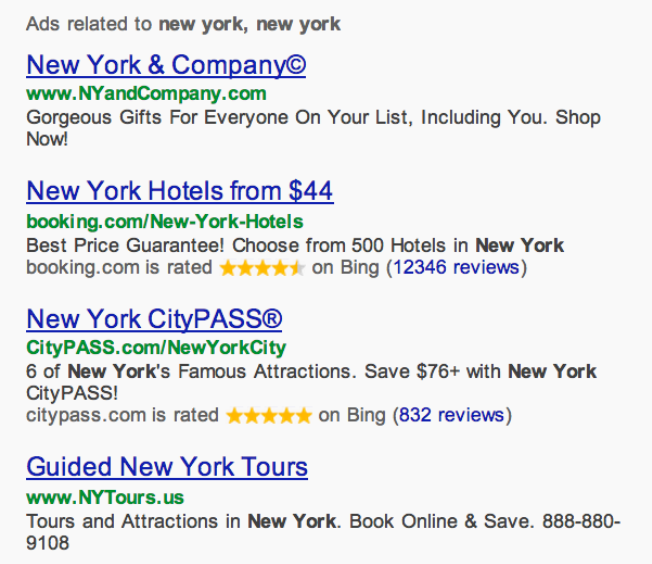

Figure 2: Sponsored Search Ads for the Keyword "new york, new york." Notice the Diversity in the Advertisers.

different intents while bidding on a particular keyword. However, due to privacy concerns and technological limitations, the extent to which many search engines allow the firms to target consumers is limited, and the majority of firms target consumers only based on keywords. This limitation may result in firms from diverse industries such as tourism, hotels, or movies bidding for the same keywords, which can lead to diverse sponsored ads as shown in Figure 2. Differences in consumers' search interests and advertisers' intents 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). The likelihood of an ad being clicked by a consumer, *ceteris paribus*, depends on the relevance of the advertiser's intent to the consumer's search interest. Any potential mismatch between the consumer's search interest and the advertiser's intent might reduce the efficiency of search advertising. However, the magnitude of the mismatch varies across keywords. Some keywords, such as "antivirus," are specific, and consumers who use these keywords share the same search interest. Consumers who search for "antivirus" and advertisers who bid on "antivirus" are likely to refer to the same type of products. Hence, consumers are likely to find all the search results (both organic search results and sponsored ads) displayed for "antivirus" relevant to their search interests. Other keywords,

such as "Mars" and "new york, new york" are ambiguous and may reflect a variety of search interests. Consumers who search for "Mars" and advertisers who bid on "Mars" may refer to different types of products. Consumers thereby may find that some of the organic search results or sponsored ads displayed for these ambiguous keywords irrelevant to their search interests.

In this paper, we define *keyword ambiguity* as the breadth of search interests as indicated by a keyword. We herein wish to understand how keyword ambiguity might affect keyword performance in search advertising. More specifically, we try to answer the following questions in our research.

1. How can we quantify keyword ambiguity?
2. Does keyword ambiguity affect keyword performance?
3. How does the effect of keyword ambiguity vary with ad position?

To answer the aforementioned questions, we use a rich data set from a major search engine to perform a cross-category analysis and examine the economic impact of keyword ambiguity on keyword performance in the context of search advertising.

Despite the importance of understanding consumers' search interests in search advertising, there exists little empirical research analyzing how keyword ambiguity affects consumer click behavior. One potential challenge is limited availability of data. A major advantage of our study is that we are able to examine a large variety of keywords across multiple product categories, whereas previous studies mostly focus on a single product category from a particular retailer (e.g., Ghose and Yang 2009; Agarwal et al. 2011; Rutz and Bucklin 2011). Our cross-category analysis helps us examine how CTR varies by industry and provide broad generalizations as well as focus on category-specific effects that advertisers can take into account while crafting their search advertising strategy. Another advantage of this data set is the presence of *all* competing ads for an impression, which helps us build a richer model of consumer click behavior and derive new insights.

Another challenge in examining the role of keyword ambiguity in consumer click behavior is to infer the different search interests associated with each keyword and quantify the ambiguity of the keyword. Due to the extensive nature of our data, we resort to novel machine learning techniques and propose an automatic method of categorizing keywords and examining keyword ambiguity based on topic models (Blei et al. 2003) 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, to measure the ambiguity of a keyword as the uncertainty in predicting consumers' search interests. Subsequently, we quantify the effect of keyword ambiguity on keyword click performance using a hierarchical Bayesian model by simultaneously modeling click-through rate (CTR) and ad position (rank).

We find that keyword ambiguity has two opposing effects on keyword click performance. First, higher keyword ambiguity is associated with higher baseline CTR (i.e., click probability at the top position). This phenomenon arises due to consumers substituting away from ambiguous organic results to sponsored ads. At the same time, higher keyword ambiguity is also associated with a faster decay in CTR with screen position (i.e., decrease in CTR from top to bottom). This result suggests that, although consumers are more likely to switch to sponsored ads for more ambiguous keywords than less ambiguous keywords, they are more likely to stop clicking on ads at lower positions, because fewer ads are relevant to the consumers' intents. Therefore, the overall effect of keyword ambiguity on CTR for sponsored ads at various positions is a combination of these two opposing effects. In other words, keyword ambiguity seems to benefit top-positioned ads but hurt lower-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 across keyword categories. For example, the click-throughs for certain categories, such as "adult" and "style," are more evenly distributed across positions as compared to categories such as "health." These results suggest that the position effect appears 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, and we believe that this is the first paper to compare keyword performance across different categories.

This paper makes the following contributions. First, while most of the prior empirical studies in search advertising have ignored the effect of keyword ambiguity in search advertising and implicitly assumed that all sponsored ads are relevant to all consumers, we extend to this stream of research by operationalizing the concept of keyword ambiguity and demonstrating how machine learning and computational linguistic tools such as topic models can be used to extract keyword ambiguity and other semantic characteristics of keywords based on large-scale analytics from unstructured text data. Second, we expand the search advertising literature by examining how keyword ambiguity affects keyword performance (i.e., CTR) in multi-category search advertising, which increases our understanding of the heterogeneity in the keyword performance through previously unobserved semantic characteristics. Third, this paper increases our understanding of the interactions between organic results and sponsored ads by showing that, for keywords with high ambiguity, consumers tend to substitute away from ambiguous organic results to sponsored ads, which leads to an increase in overall CTR for ambiguous keywords. Finally, our CTR model outperforms (in terms of predictive power) alternative models that do not account for keyword ambiguity and consumer heterogeneity in search interests.

This study provides implications for both advertisers and search engines. First, the approach proposed in this paper to extract keyword semantic characteristics presents a new metric, i.e., keyword ambiguity, along with other semantic characteristics of keywords, that can be used by advertisers to improve the prediction of keyword performance, evaluate keywords and craft their bidding strategy. Second, the availability of data for keywords across multiple advertisers and categories allows us to gain insights into how CTR varies by industry and ad positions. For example, our results indicate health-related keywords tend to attract higher CTR than car-related keywords at the top position; however, the CTR of health-related keywords tend to

decrease faster with positions than car-related keywords. Finally, this study provides implications for search engines to design keyword planning tools to aid advertisers when choosing potential keywords and to improve the quality of sponsored ads served in response to a consumer search query.

LITERATURE REVIEW

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

Search Advertising

During the past decade, the increasing popularity of search advertising has motivated research from multiple areas. Theoretical research on search advertising includes Edelman et al. (2007), Varian (2007), Athey and Ellison (2011), and Weber and Zheng (2007), among other papers. Most of these studies focus on auction design and bidding strategies of firms. Empirical research on search advertising is also growing rapidly (e.g., Ghose and Yang 2009; Animesh et al. 2010; Yang and Ghose 2010; Agarwal et al. 2011; Animesh et al. 2011; Agarwal et al. 2015). Online Appendix A provides an overview of previous empirical studies on search advertising. Most of these studies primarily use keyword-level aggregate data provided by advertisers in particular industries to study ad performance. Exceptions are Jerath et al. (2014), Jeziorski and Segal (2015), and Yao and Mela (2011), who use individual-level data provided by search engines.

In particular, our study is closely related to prior work that examines the impact of keyword characteristics on keyword performance (e.g., Ghose and Yang 2009; Yang and Ghose 2010; Agarwal et al. 2011; Rutz and Bucklin 2011). In these studies, the keyword characteristics are typically hand-coded on a small scale. For example, Rutz et al. (2011) 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 automatically characterize the topical content and category of each keyword 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 this paper, we use a data set that contains consumer click-through information for a large number of keywords across multiple product categories from different types of advertisers, which would have been extremely difficult without machine learning techniques. By analyzing the characteristics extracted automatically for each keyword, we are able to examine click performance of keywords across multiple product categories.

One of the gaps in the existing literature is the interplay between organic and sponsored search results (Jansen and Resnick 2006; Jansen et al. 2007; Buscher et al. 2010; Danescu-Niculescu-Mizil et al. 2010; Agarwal et al. 2015; Yin et al. 2014). Earlier experimental studies have shown that most consumers examine organic results first before sponsored ads (Jansen and Resnick 2006; Jansen et al. 2007), but the relevance ratings for organic results and those for sponsored ads are practically the same (Jansen 2007). Danescu-Niculescu-Mizil et al. (2010) analyze a click-through data set and find that the relationship between organic and sponsored search results depends on whether the keyword is navigational or non-navigational.[2] In particular, they find a substitution effect between the most clicked organic result and the most clicked sponsored ad when the keyword is navigational, and a complementary effect when the keyword is non-navigational. A few studies have found evidence of a substitution effect between organic results and sponsored ads. An eye tracking study by Buscher et al. (2010) finds that, when a list of less relevant ads is displayed, the organic results receive more visual attention. Using data from a commercial search engine, Yin et al. (2014) examine the correlation between the CTR of organic results and the CTR of sponsored ads and find a negative correlation between the two types of search results. In addition, Agarwal et al. (2015) analyze click-through and conversion data from an online retailer using a hierarchical Bayesian

[2]Navigational keywords as keywords used by users when "the immediate intent is to reach a particular site" (Broder 2002).

model and find that the competitive intensity in the organic listing has a negative impact on the click performance of sponsored ads, suggesting a substitution effect between the organic listing and sponsored listing. Although these studies provide an initial investigation into the relationship between organic results and sponsored ads, we aim to infer the role of keyword ambiguity in the interplay between organic and sponsored search results.

Machine Learning and Text Mining

A stream of research has recently emerged in information systems and related areas that applies machine learning and text mining techniques in examining online markets (e.g., Gu et al. 2007; Archak et al. 2011; Ghose et al. 2012; Lee et al. 2016). Gu et al. (2007) apply text mining to assess the quality of postings in virtual communities to examine users' valuation of virtual communities and the differentiation among virtual communities. Ghose and Ipeirotis (2011) and Archak et al. (2011) use text mining to extract multiple aspects of online review texts to identify text-based features and study their impact on review helpfulness and product sales, respectively. Netzer et al. (2012) combine text mining and semantic network analysis to understand the brand associative network and the implied market structure. Ghose et al. (2012) use text mining and image classification to analyze the economic effects of user-generated content and crowd-sourced content, and design a utility-based ranking system for products that can lead to an increase in consumer surplus. In our study, we propose to apply 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. Applications of topic models in information systems research include the analysis of blog content (Singh et al. 2014), stock recommendation messages (Aral et al. 2011), and firms' financial reports (Bao and Datta 2014).

Semantic Ambiguity

Prior psycholinguistic studies have documented the effect of semantic word ambiguity on language processing and comprehension (e.g., Rayner and Duffy 1986; Kellas et al. 1988; Borowsky and Masson 1996; Rodd et al. 2002; Hoffman and Woollams 2015). However, to the

best of our knowledge, there is no empirical study that focuses on the ambiguity of keywords and examines the economic impact of keyword ambiguity on ad performance in the context of search advertising.

From a methodology perspective, studies in this stream of literature measure word ambiguity based on the number of distinct meanings of a word as listed in dictionaries, or as evaluated by human coders (Jastrzembski 1981; Kellas et al. 1988; Borowsky and Masson 1996; Rodd et al. 2002, 2004). There are a few other corpus-based measures of word ambiguity proposed in prior studies. For example, Adelman et al. (2006) measure contextual diversity, i.e., the number of contexts in which a word occurs, by counting the number of documents in a corpus that contain a given word. McDonald and Shillcock (2001) measure contextual distinctiveness of a word based on the informativeness about the word's context of use. A few recent studies (Hoffman et al. 2013; Hoffman and Woollams 2015) take a different approach based on latent semantic analysis (LSA; Landauer and Dumais 1997) that uses word co-occurrence in a corpus to construct distinct contexts, where each word is subsequently represented as a vector. They then compute the word's entropy in the corpus to measure the semantic diversity of a word.

However, the above approaches used in prior literature all focus on measuring the ambiguity of individual words and are less feasible for search keywords, because search keywords are often phrases that tend to have a high level of complexity for the following reasons. First, a dictionary or encyclopedia is likely to have limited coverage for brand or firm names. Second, consumers may use slightly different forms of keywords when searching for the same brand. For example, "verizon" "verizon wireless" and "verizon mobile" all refer to the same brand, and some forms may not be included in dictionaries or encyclopedia such as Wikipedia. Third, a large portion of the keywords, such as "free PC poker games," are phrases that are not covered by a dictionary or encyclopedia. Our approach to measure keyword ambiguity applies a topic model on keyword-specific search results and subsequently computes the entropy (i.e., diversity) of the topic distribution of a keyword. This approach is easily scalable and can work for any keyword used by users on search engines.

Agarwal et al. (2011) develop a related keyword characteristic called specificity, which they define as "the level in the product hierarchy of the advertiser." The key distinction between keyword ambiguity and specificity is that keyword ambiguity measures the breadth of topics, while specificity measures the depth within a topic when only one advertiser is examined. For example, based on the definition of specificity by Agarwal et al. (2011), "formal blue shirt" and "Levi's shirt" have higher specificity than "shirt," although they are all related to the clothing topic. However, based on our definition of keyword ambiguity, all these keywords are all most likely related to the topic "clothing," thus have a similar level of ambiguity.

THEORETICAL BACKGROUND

A consumer uses search engines to find websites that meet her need. When a consumer searches using a keyword, she sees a list of organic search results on the search result page along with a list of sponsored ads. Typically, she explores the organic search results before sponsored ads (Jansen and Resnick 2006; Jansen et al. 2007; Jerath et al. 2014). If the organic search results satisfy her need, she may not explore the sponsored ads at all. However, when she explores the sponsored ads, she evaluates them in a sequential manner by focusing her attention on the top-positioned ad first, and then evaluates each ad from top to bottom (Granka et al. 2004).[3] Before clicking on an ad, the consumer does not have complete information about the quality of each ad, but she can read the ad description on the search result page and infer its relevance. When deciding whether to click on each ad, the consumer takes into account the expected utility that she obtains from clicking on the ad based on the ad description and screen position, as well as the cost of evaluating the ad (i.e., search cost). After clicking on an ad, she may continue the search by evaluating another ad or abandon the search session entirely (Agarwal and Mukhopadhyay 2016). Due to the non-negligible search costs of evaluation, fewer consumers end up visiting lower positions (Arbatskaya 2007; Agarwal and Mukhopadhyay

[3]Several behavioral studies have provided strong evidence of such sequential search process (e.g., Saad and Russo 1996; Moorthy et al. 1997). This assumption is especially widely used in the online search environment by both empirical studies (e.g., Ghose and Yang 2009; Animesh et al. 2010, 2011; Agarwal et al. 2015; Jeziorski and Segal 2015; Agarwal and Mukhopadhyay 2016) and analytical studies (e.g., Athey and Ellison 2011; Chen and He 2011) in search advertising.

2016). Previous empirical literature has also established that the click performance often decays with ad position (e.g., Ghose and Yang 2009; Agarwal et al. 2011; Rutz et al. 2011; Yao and Mela 2011; Rutz et al. 2012).

When a given keyword is ambiguous and associated with heterogeneous search interests, search engines are likely to offer a diverse list of organic results as well as possibly a diverse set of sponsored ads to serve different needs. In this case, the consumer might find that only a limited number of organic results are relevant to her search intent. On the one hand, there may be a complementary effect between the organic results and sponsored ads. In particular, based on the diverse organic results, the consumer might infer that the sponsored ads are also diverse and less relevant, hence will be less likely to click on them (Jeziorski and Segal 2015; Athey and Ellison 2011). As a result, keyword ambiguity might decrease the consumer's propensity of clicking on sponsored ads. On the other hand, there may also be a substitution effect between the organic results and sponsored ads, as the two types of search results compete with each other for consumer attention (Agarwal et al. 2015). With a diverse list of organic results, it is possible that the top organic results might not fulfill the consumer's needs, hence she may then switch to sponsored ads to look for relevant information (Jansen et al. 2007).[4] A few studies have examined the correlation between the CTR of organic results and the CTR of sponsored ads descriptively, and found such substitution effect. For example, Yin et al. (2014) use an archival data set from a commercial search engine, and find a negative correlation between the organic results and the sponsored ads placed above organic results. As a result, keyword ambiguity (with a diverse list of organic results) might increase consumers' propensity of clicking on sponsored ads.

Therefore, it is unclear whether a diverse list of organic results associated with higher keyword ambiguity would turn consumers away from sponsored ads, thus reducing their propensity of

[4]In the case of an unambiguous keyword, consumers might be presented with a list of organic search results that can fully satisfy their search need. Due to limited time and attention to explore the ads (Agarwal et al. 2015) and information satiation (Jeziorski and Segal 2015), they might be less likely to click on sponsored ads under this condition.

clicking on sponsored ads, or draw consumers to sponsored ads, thus increasing their propensity to click on sponsored ads. Although we do not directly have data on how consumers respond to organic results, our analysis on sponsored ads can help us better understand how the diversity of organic results affect consumers' propensity to click on sponsored ads.

Even when the consumer starts interacting with the sponsored ads, it is not clear how keyword ambiguity might affect the depth of search (e.g., the decay in the CTR with position). As an ambiguous keyword might be associated with a diverse list of sponsored ads, fewer ads may be relevant to a consumer's search intent. Since the consumer can identify the relevance of an ad (with some level of certainty) based on the ad description on the search result page (Jansen 2007), she will click on fewer ads for ambiguous keywords as compared to ads for unambiguous keywords. This phenomenon will results in a faster decay in CTR with position.

In addition, consumer heterogeneity might play an important role in the impact of keyword ambiguity on keyword performance. Economic search theory suggests that an increase in search cost reduces search intensity (Weitzman 1979). If consumers with lower search cost use more ambiguous keywords, we will observe that higher keyword ambiguity is associated with a higher likelihood to engage with sponsored ads as well as a smaller decay in CTR with position (as such consumers with low search costs will naturally search and click more intensively). On the other hand, if consumers with higher search costs use more ambiguous keywords, then the opposite results will be observed (i.e., higher keyword ambiguity will be associated with a lower likelihood to engage with sponsored ads as well as a faster decay in CTR with position).

Therefore, the goal of this paper is to empirically examine how keyword ambiguity might affect the click performance (both the baseline CTR and the decay rate) of sponsored ads given these aforementioned potential factors. In the empirical analysis, we discuss the *net result* of these factors and try to identify which mechanisms might be dominant in this context.

DATA

The data set 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 data set comprises the keyword (a word or a phrase of more than one word) a consumer searched and a list of sponsored ads shown to the consumer. The maximum number of ads shown per impression is eight.[5] For each ad displayed, we observe whether it was clicked during an impression. Note that although we have a unique ad identifier and thus can track an ad across impressions, we do not have any ad-specific information. Since our data set is derived from the search engine, we do not have post-click information (e.g., conversions), unlike some previous papers (Ghose and Yang 2009; Agarwal et al. 2011) that 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 follow prior literature (e.g., Rutz et al. 2011; Rutz and Trusov 2011; Jerath et al. 2014; Agarwal et al. 2015) and remove the low-performing keywords for three specific reasons. Firstly, these keywords are not as relevant to advertisers, as ads associated with these keywords never get clicked in our sampling period (Rutz et al. 2011). Secondly, low performing keywords tend to have relatively low search volume and result in sparse data. Although the empirical model we use in the paper (i.e., a hierarchical Bayesian model) can deal with low click-through rates, the inclusion of very sparse data in the estimation may affect the ability to recover the heterogeneous parameters and may lead to inferior model performance (Rutz and Trusov 2011). Thirdly, given the complexity of our hierarchical Bayesian model, estimation based on all the keywords would take a substantial amount of time. Hence, we restrict our analysis to a subset of keywords. We choose to ignore keywords containing domain names, because users who use these keywords know exactly which websites they wish to visit, and these keywords are unlikely

[5]We do not have information on organic search results. Although we do not have individual identifiers, which may restrict our ability to track individuals over time, every impression in our data set has a unique identifier.

to lead to additional traffic for the websites.

The full data set includes 12,790 distinct keywords from more than 4.6 million impressions. More than 0.17 million unique ads are displayed, 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. (2014) that very few searches lead to clicks on sponsored ads because user needs might be met by organic results.

Table 1, Part (A) presents summary statistics for the full data set. *CLICK* is an indicator of whether the ad is clicked. *POS* denotes the ad position in an impression, ranging from 0 to 7. *NUM_ADS* denotes the number of ads during an impression, which measures the competitive intensity for a keyword. *NUM_WORDS* denotes the number of words in the keyword. *LOG_IMP* denotes the natural log of the total number of times consumers search for a particular keyword in the data set, which is analogous to the popularity measure used in Jerath et al. (2014). In the next section, we discuss how to use machine learning tools to extract additional keyword semantic features, such as keyword ambiguity, the presence of brand and location names, and transactional intent. Online Appendix F reports the correlation matrix among variables.

Our unique, large data set allows us to provide insights that can be generalized across multiple categories. This cross-category analysis can also help us identify differences between keywords across different categories. Online Appendix A presents a comparison between the data set used in our paper and prior empirical research in search advertising. Note that the full data set is prohibitively large for estimation. Therefore, we use all 12,790 keywords (referred to as the full data set) for extracting semantic characteristics, and randomly sample 10,000 impressions (referred to as the focal data set) for subsequent empirical analysis on click-through performance. The focal data set used for estimation includes 2,625 unique keywords and 10,750 unique ads, resulting in 47,403 ads displayed. A comparison of the focal data set with the full

data set as shown in Table 1 indicates that the focal data set is a fairly representative sample of the full data set.

Table 1: Summary Statistics

Variable	(A) Full Data Set					(B) Focal Data Set				
	Obs	Mean	SD	Min	Max	Obs	Mean	SD	Min	Max
Impression-ad level										
CLICK	24,149,179	0.03	0.16	0	1	47,403	0.02	0.14	0	1
POS	24,149,179	2.66	2.08	0	7	47,403	2.50	2.04	0	7
Impression Level										
NUM_ADS	4,641,738	5.19	2.42	1	8	10,000	4.74	2.45	1	8
Keyword Level										
NUM_WORDS	12,790	1.74	0.68	1	5	2,625	1.76	0.69	1	4
LOG_IMP	12,790	4.87	0.99	0.69	12.59	2,625	5.87	1.31	3.40	12.59

EXTRACTING SEMANTIC FEATURES USING MACHINE LEARNING

In this section, we demonstrate how machine learning and computational linguistic tools such as topic models can be used to extract keyword topics, keyword ambiguity, and other semantic characteristics of keywords based on unstructured text data. To the best of our knowledge, we are not aware of any approach that aims to measure the ambiguity of search keywords. Since keywords are words or phrases with a high level of complexity, it is difficult to extract the contextual meanings of a large number of keywords either manually or using a dictionary approach. Given the large number of keywords used in our analysis, we resort to unsupervised machine learning methods. Specifically, we apply a topic model on a corpus constructed using keyword-specific search results, and subsequently, compute the entropy of the topic distribution of each keyword to quantify keyword ambiguity. Our machine learning approach is easily scalable and can work for a large number of keywords.

A Generative Model of the Topical Content of Keywords

The major challenge in examining the impact of keyword ambiguity on click performance is to quantify keyword ambiguity. We model the ambiguity of each keyword based on probabilistic topic models from machine learning and natural language processing (Blei et al. 2003). 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. Topic models have been applied to many contexts, such as the analysis of blog content (Singh et al. 2014) and stock recommendation messages (Aral et al. 2011). It is an unsupervised learning approach in that we do not know the topics ex ante. The use of a supervised learning approach with predefined topics may significantly restrict the discovery of hidden interesting topics, especially for those topics that are less popular (which we may not consider as a "label" for classification in a typical supervised learning process).

Corpus Construction and Document Pre-processing

We first construct a corpus of documents that describe the information content conveyed by the keywords. As keywords are usually words or short phrases, obtaining the true contextual meanings of a keyword based on the keyword itself is usually difficult. To solve this challenge, we use Google organic search results to augment the keyword data set and better understand the semantic meanings associated with each keyword, because Google organic search results generated based on the classical theory of document relevancy provide a reasonable approximation. This approach of using organic search results to enrich keyword meanings has been used in prior studies in information retrieval (e.g., Broder 2002; Dai et al. 2006; Abhishek and Hosanagar 2007).

For each keyword in our full data set, we extract the title and textual content of the brief description from each of the top-50-ranked Google organic search results[6] (Figure 1), to

[6]We collected the organic search results using Google's search API where no personalization is implemented. Therefore, search engine results personalization is unlikely to affect our measurement of keyword ambiguity.

construct the corresponding keyword-specific document.[7] 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 documents following a standard procedure (e.g., Aral et al. 2011). We first remove annotations and tokenize the sentence into distinct terms, and then remove stop words using a standard dictionary.

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. Recall that in the previous subsection, we construct a document for each keyword that best reflects the contextual information of the keyword. We then apply the LDA model to the corpus of documents to infer the topics associated with each document. In particular, 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

[7]The Google organic search results were first collected in 2013. We have conducted several robustness checks. First, we repeated our analysis based on different numbers of Google organic search results (i.e., top-60, top-80 and top-100). The results are robust to the number of organic search results used to construct the corpus for topic modeling. The comparison is presented subsequently as a robustness check.

Second, we re-collected Google organic search results for the focal data set (2,625 keywords) in 2016. We also tried to create a proxy for Google organic search results in 2007 by limiting the date range of results to only those before August 9, 2007 (by the time our click-through data were collected) to obtain new topic entropy values for comparison. The computed topic entropy based on different years of organic search results are highly correlated, suggesting that entropy values seem to be fairly robust to organic search results collected from different years.

model with a different number of topics (20, 50, and 100). The details of the LDA model are provided in Online Appendix B.

The most frequent words identified for the 20-topic model based on the topic probability of each word are presented in Figure 3, where topics are color coded. For convenience, we assign a label to each topic (e.g., "sport," "music," and "food") based on its high-frequency words. For example, documents related to the topic "style" often contain words such as "dress," "party," "woman," and "fashion." In addition, words that occur frequently in multiple topics are highlighted in brown, such as "free," "shop," and "find." We also present the topic distributions estimated from the 20-topic model for a sample of keywords in Online Appendix C.

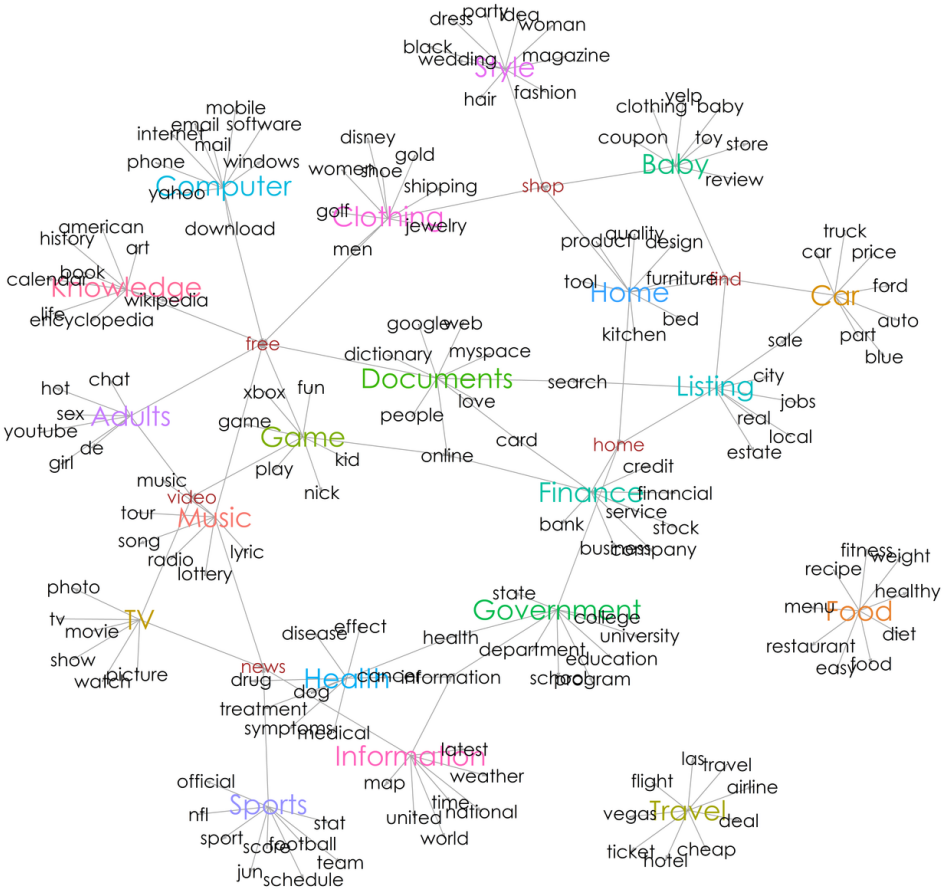

Figure 3: Frequent Words in Each Topic

Topic Entropy as a Measure of Keyword Ambiguity

We propose using *topic entropy* to measure the ambiguity of a keyword, which captures the uncertainty of a keyword/document's topic distribution (Hall et al. 2008). 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. 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, as obtained from the LDA model. We then define the topic entropy of keyword k as follows:

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

where T is the total number of topics.

Figure 4 illustrates the posterior topic probabilities and topic entropy for two sample keywords, "free anti virus" and "express." For the keyword "free anti virus," the estimated probability that it is related to the topic "computer" is extremely high (0.93), and low for other topics, suggesting that "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 flat 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.[8] The high correlations among entropy values derived based on different numbers of topics also suggest that entropy seems to be fairly robust to the number of topics specified in the LDA model.

[8]The number of topics T is pre-specified before estimating the LDA model. As can be seen in Table 2, the maximum entropy value depends on the number of topics chosen. Entropy for keyword k is the smallest when there exists $t \in \{1,...,T\}$ such that $\hat{\theta}_{kt} = 1$; Entropy is the largest when for all $t \in \{1,...,T\}$, $\hat{\theta}_{kt} = 1/T$. Therefore, with T topics, entropy ranges from 0 to $log(T)$.

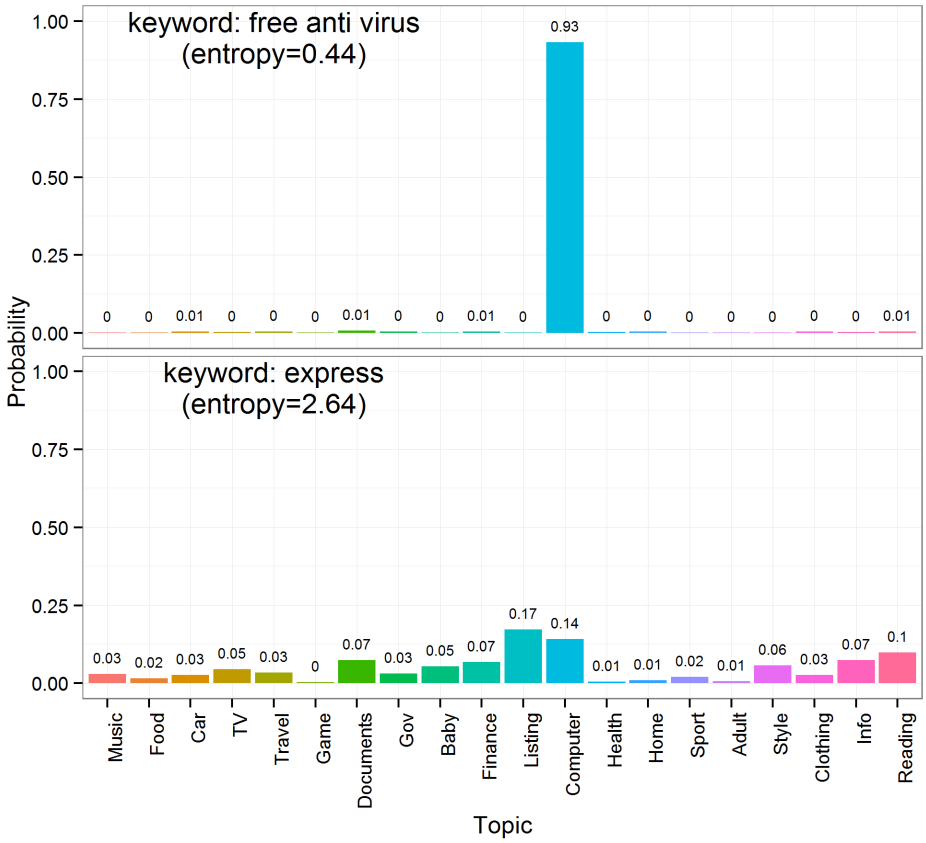

Figure 4: Topic Distribution and Topic Entropy: A Demonstration

Extracting Other Keyword Semantic Features

Previous studies use brand name (e.g., Ghose and Yang 2009; Yang and Ghose 2010; Jerath et al. 2014) 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) that matches each keyword against a list of brand names and locations. The detailed procedures are available in Online Appendix D.

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; Goldenberg et al. 2012). Consumers may have different search intents, and the keywords they use in their search

Table 2: Summary Statistics for Topic Entropy

	Mean	SD	Min	Max	Correlation	
					20 Topics	50 Topics
20 Topics	1.60	0.45	0.338	2.99		
50 Topics	2.02	0.52	0.43	3.90	0.88	
100 Topics	2.29	0.55	0.49	4.58	0.86	0.91

activities may reflect their intents. Broder (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. We propose to infer the transactional intent of each keyword using the keyword's corresponding Google organic results. To do so, we first compose a list of transactional words based on Dai et al. (2006) and general knowledge. These transactional words are listed in Online Appendix E. Then, for each keyword, we count the frequency of transactional words in the corresponding Google organic results of the keyword. Lastly, we use *LOG_TRANS*, the natural log of the frequency of transactional words, to measure the keyword's transactional intent.

Summary Statistics of Extracted Features

We present summary statistics of extracted semantic features in Table 3. *TOPIC_ENTROPY* measures keyword ambiguity based on the results of a topic model with 20 topics. *BRAND* is a dummy variable indicating whether the keyword contains brand names. *LOCATION* is a dummy variable indicating whether the keyword contains city or state names. In our data set, 17% of the keywords are classified as containing brand information, and 13% are classified as containing location information. *LOG_TRANS* measures a keyword's transactional intent. We present the correlation matrix among all variables in Online Appendix F.

Table 3: Summary of Extracted Features

Variable	(A) Full Data Set				(B) Focal Data Set			
	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
TOPIC_ENTROPY	1.60	0.45	0.34	2.99	1.59	0.45	0.34	2.69
BRAND	0.17	0.38	0	1	0.24	0.42	0	1
LOCATION	0.13	0.34	0	1	0.11	0.31	0	1
LOG_TRANS	2.17	1.08	0	5.51	2.16	1.09	0	5.31

A MODEL OF CLICK-THROUGH AND AD POSITION

We build an ad-impression level model to capture how CTR varies with ad position. Estimating any ad-specific model would take an extremely large amount of time. Hence, we choose to perform our analysis based on a random sample of 10,000 impressions instead of using the full data set.[9] We begin this discussion by formally introducing a hierarchical Bayesian model for keyword performance. Hierarchical Bayesian models have been widely used in search advertising literature (e.g., Ghose and Yang 2009; Yang and Ghose 2010; Agarwal et al. 2011, 2015). We propose to use a hierarchical Bayesian model that allows for heterogeneity in keyword performance at several levels: topic level, keyword level, and ad level. To control for the endogeneity of ad position, we jointly model CTR and ad position by correlating the error terms of the click-through equation and the position equation. Our main model is nonlinear in that the click-through rate is modeled using a probit model. The model we propose is more suitable in our case than standard random utility models in that our model allows for not only heterogeneity, but also position endogeneity and nonlinearity.

Modeling CTR

We assume that a consumer's decision of whether to click a sponsored ad is a binary choice that depends on ad characteristics, keyword characteristics, impression-level characteristics, and unobserved impression-ad-level characteristics. Our unit of analysis is an ad-impression.

[9] We experimented with different random samples, and the results are qualitatively similar.

Specifically, we model that the utility that a consumer derives from clicking on ad a in impression i using keyword k_i that belongs to topic t as follows:[10]

$$U_{iat} = \beta_{0,k_i,t} + \beta_{1,k_i,t} POS_{ia} + \beta_2 NUM\_AD_i + \tau_a + \eta_{ia}. \quad (2)$$

The consumer clicks on the ad when $U_{iat} \geq 0$. In Equation (2), $\beta_{0,k,t}$ is an intercept term that varies by keyword and topic, which measures the baseline CTR associated with ads for keyword k and topic t. In other words, as $\beta_{0,k,t}$ increases, ads at all positions are more likely to get clicked. On the other hand, $\beta_{1,k,t}$ represents the effect of position on CTR. We expect that $\beta_{1,k,t}$ is negative, and larger $\beta_{1,k,t}$ indicates a smaller decay rate with position. The intercept term $\beta_{0,k,t}$ and the coefficient of position $\beta_{1,k,t}$ might depend on the characteristics of keyword k. In particular, to examine the effect of keyword ambiguity on CTR, we assume that $\beta_{0,k,t}$ is a function of keyword ambiguity (*TOPIC_ENTROPY*), and other keyword characteristics including *NUM_WORDS*, *BRAND*, *LOCATION*, *LOG_TRANS*, and *LOG_IMP* that serve as controls in our analysis. Similarly, to examine how keyword ambiguity moderates the effect of ad position on CTR, we assume that $\beta_{1,k,t}$ is also a function of keyword ambiguity and other keyword characteristics. This hierarchical approach we use to capture the effect of keyword characteristics is similar to the ones used by Ghose and Yang (2009), Yang and Ghose (2010), and Agarwal et al. (2011). The parameter β_2 captures how the number of competing ads in an impression affects CTR. τ_a captures unobserved ad specific factors such as ad quality, and η_{ia} is an impression-ad specific error term.

To capture how the baseline CTR, $\beta_{0,k,t}$, varies by keyword and topic, we model $\beta_{0,k,t}$ as follows:

$$\beta_{0,k,t} = \gamma_{0,t} + X_k' \Delta_0^{\beta} + u_{0,k}^{\beta}, \quad (3)$$

where $\gamma_{0,t}$ captures the heterogeneity in the baseline CTR across topics, and $u_{0,k}^{\beta}$ is the idiosyncratic error term. X_k is a vector of keyword characteristics including *TOPIC_ENTROPY*, *NUM_WORDS*, *BRAND*, *LOCATION*, *LOG_TRANS*, and *LOG_IMP*. Δ_0^{β} is a vector of coefficients

[10] Here we use k_i to denote the keyword associated with impression i. In the following discussion, we may use k to denote the focal keyword without reference to a specific impression.

that capture the effects of these keyword characteristics on the baseline CTR. Specifically, the effect of keyword ambiguity on the baseline CTR is captured by $\Delta^{\beta}_{0,TOPIC\_ENTROPY}$.

The coefficient of position, $\beta_{1,k,t}$, is modeled as follows:

$$\beta_{1,k,t} = \gamma_{1,t} + X_k' \Delta^{\beta}_1 + u^{\beta}_{1,k}, \tag{4}$$

where $\gamma_{1,t}$ captures the topic specific effect of position, and $u^{\beta}_{1,k}$ is the idiosyncratic error term. Δ^{β}_1 is a vector of coefficients that capture the effects of the keyword characteristics on the decay with position, $\beta_{1,k,t}$. Specifically, the effect of keyword ambiguity on the decay in CTR with position is captured by $\Delta^{\beta}_{1,TOPIC\_ENTROPY}$.

Search engines rank ads based on many ad specific factors including quality and relevance, both of which are unobserved and potentially correlated with ad position. Therefore, we include τ_a in Equation (2) to capture unobserved ad specific factors such as ad quality. We assume that τ_a follows a normal distribution $N(0, v^{\tau})$.[11] We assume that $\gamma_t = (\gamma_{0,t}, \gamma_{1,t})'$ follows a multivariate normal distribution,

$$\gamma_t \sim MVN(0, \Psi), \tag{5}$$

and the error vector $u^{\beta}_k = (u^{\beta}_{0,k}, u^{\beta}_{1,k})'$ follows a multivariate normal distribution $u^{\beta}_k \sim MVN(0, \Omega^{\beta})$.

Modeling Ad Position

Most major search engines determine ad positions based on factors such as advertisers' past performance and ad relevance. Therefore, advertising strategies are often endogenous, which makes it challenging to examine the causal impact of sponsored ads.

In Equation (2), we use τ_a to control for unobserved ad specific heterogeneity such as ad quality. However, other unobserved factors such as ad relevance that may vary across impressions cannot be simply captured by τ_a. For example, search engines may position more

[11] In the next subsection, we discuss how we use a Hausman style instrumental variable (Hausman 1996) to control for position endogeneity caused by ad-keyword specific factors such as ad relevance.

relevant ads first based on the keyword used by the consumer. As a result, ad relevance may be an unobserved, ad-keyword specific factor, which may affect both ad position and CTR. If we do not address the endogeneity of ad position, our estimates may be biased. To alleviate the concern of endogenous ad position, some prior studies have designed field experiments (e.g., Agarwal et al. 2011; Animesh et al. 2011) to measure the causal effect of ad position in the context of sponsored ads. However, experimentation is infeasible in our context because of the cross-category nature of our data. Other studies leverage information on bidding to address the endogeneity concern (e.g., Narayanan and Kalyanam 2015; Yao and Mela 2011).

To account for the potential endogeneity of ad position, we use a Hausman Instrumental Variable (IV) approach (Hausman 1996) that has been commonly used in economics (Nevo 2000, 2001) and related fields (Che et al. 2007; Ghose et al. 2012). Specifically, for ad a in impression i when keyword k_i is searched, let $\overline{POS}_{a,-k_i}$ denote the average position of ad a displayed for all keywords other than keyword k_i. To use $\overline{POS}_{a,-k_i}$ as an instrumental variable for POS_{ia}, $\overline{POS}_{a,-k_i}$ should be correlated with POS_{ia}, conditional on other covariates, but uncorrelated with the error term η_{ia} in Equation (2). One potential concern is that ad relevance may be a potential "omitted" variable in η_{ia}. Since ad relevance is ad-keyword specific, it is unlikely that $\overline{POS}_{a,-k_i}$, the average position of ad a for keywords other than k_i, is correlated with the relevance of ad a to keyword k_i. Therefore, $\overline{POS}_{a,-k_i}$ may serve as an instrumental variable for POS_{ia} because it is correlated with POS_{ia} and uncorrelated with the error term η_{ia}. In our context, supply-side factors such as advertisers' willingness to bid and advertising budget are correlated across keywords and affect ad positions (Athey and Ellison 2011), thus the positions of the same ad across keywords should be correlated. However, these supply-side factors are unlikely to affect users' click-through behavior directly. Therefore, the average position for other keywords can be used as an instrumental variable for ad position.

We follow Ghose and Yang (2009) and Agarwal et al. (2011) by simultaneously modeling click-through rate and ad position. With $\overline{POS}_{a,-k_i}$ as an instrumental variable for POS_{ia}, we

model ad position as

$$POS_{ia} = \phi_{0,k_i} + \phi_{1,k_i}\overline{POS}_{a,-k_i} + \phi_2 NUM\_AD_i + \varepsilon_{ia}. \tag{6}$$

To capture the impact of keyword characteristics on $\phi_{0,k}$ and $\phi_{1,k}$, we assume that

$$\phi_{0,k} = X_k'\Delta_0^\phi + u_{0,k}^\phi, \tag{7}$$

and

$$\phi_{1,k} = X_k'\Delta_1^\phi + u_{1,k}^\phi, \tag{8}$$

where Δ_0^ϕ and Δ_1^ϕ capture the effects of keyword characteristics on $\phi_{0,k}$ and $\phi_{1,k}$, respectively. $u_k^\phi = (u_{0,k}^\phi, u_{1,k}^\phi)' : 2 \times 1$ is a vector of error terms following a multivariate normal distribution $MVN(0, \Omega^\phi)$.

To capture the endogeneity of ad position, we allow η_{ia}, the error term from the CTR equation (Equation (2)), and ε_{ia}, the error term from the position equation (Equation (6)), to be correlated. We assume that $(\eta_{ia}, \varepsilon_{ia})'$ follows a multivariate normal distribution:

$$(\eta_{ia}, \varepsilon_{ia})' = MVN(0, \Lambda), \quad \Lambda = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma^2 + \sigma_{12}^2 \end{pmatrix}, \tag{9}$$

where Λ is a 2×2 covariance matrix capturing the correlation between CTR and ad position. For identification, the variance of η_{ia} is normalized to 1. Following Kai (1998), we use a parametrization in Equation (9) by assuming $\sigma^2 = var(\varepsilon_{ia}|\eta_{ia}) = \Lambda_{22} - \sigma_{12}^2$ to simplify our estimation.

Likelihood Function and Topic Proportions

In the previous section, we describe a model of sponsored ad performance conditional on the topic membership of the keyword. The likelihood of observing the data for a keyword conditional on the topic is given by Equations (2) and (6). However, as we have discussed earlier, each keyword has a distribution over a list of topics. Taking into account the topic distribution of each keyword, we obtain the following unconditional probability of observing the

data:

$$L(\text{Data}|X,\hat{\theta};\Theta)=\prod_i f(CLICK_i,POS_i|X_i,\hat{\theta}_{k_i};\Theta)=\prod_i\sum_t \hat{\theta}_{k_i,t} f(CLICK_i,POS_i|X_i,t;\Theta),$$

where Θ denotes all parameters to be estimated from the model. Here, $\hat{\theta}_{kt}$ represents the probability that keyword k belongs to topic t, as estimated in the LDA model. 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 earlier.

EMPIRICAL ESTIMATION AND RESULTS

Model Estimation

We estimate our hierarchical Bayesian model using Markov Chain Monte Carlo (MCMC) techniques (Gelman et al. 2003). In particular, we use a Gibbs sampling procedure, where parameters are sampled iteratively from their conditional distribution given the data and all other parameters.[12] Details of the Gibbs sampling procedure are presented in Online Appendix G. We mean-center all keyword characteristics so that $\gamma_{0,t}$ and $\gamma_{1,t}$ can be interpreted as estimates of $\beta_{0,k,t}$ and $\beta_{1,k,t}$ for a typical keyword in topic t of which the covariates are set to mean values. We use a sample of 10,000 impressions (the focal data set) as the estimation sample. Based on the estimation sample, we run two MCMC chains, each with 50,000 iterations. We discard the first 25,000 iterations of each chain, and mix the last 25,000 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 values estimated from a 20-topic LDA model for estimating the hierarchical Bayesian model.[13]

[12]Note that we model CTR conditional on the topic related to keyword k_i used in impression i. As the topic is not observed, we use a data augmentation approach by simulating topic assignment based on membership probabilities $\hat{\theta}_{k_i,t}$, where $\hat{\theta}_{k_i,t}$ is estimated from a topic model. In addition, U_{iat} is also a latent variable that involves data augmentation.

[13]We 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.

Estimation Results

Estimates for CTR

The estimated coefficients of different variables on the baseline CTR ($\beta_{0,k,t}$) are presented in Column 1 of Table 4. The effect of keyword ambiguity (*TOPIC_ENTROPY*) on $\beta_{0,k,t}$ is positive and statistically significant, indicating that higher keyword ambiguity is associated with higher baseline CTR. When consumers search using an ambiguous keyword, the search engine is likely to return a diverse list of organic search results. This might cause consumers to be dissatisfied with the organic search results, and they may turn to sponsored ads and look for an alternative that meets their needs.

Table 4: Estimation Results for CTR

	(1) Baseline ($\beta_{0,k,t}$)		(2) *POS* ($\beta_{1,k,t}$)		(3) *NUM_ADS*(β_2)	
Intercept	-2.308***	(0.122)	-0.726***	(0.045)	0.166***	(0.017)
TOPIC_ENTROPY	0.192***	(0.069)	-0.134***	(0.049)		
NUM_WORDS	0.040	(0.045)	-0.035	(0.032)		
BRAND	0.033	(0.064)	-0.050	(0.045)		
LOCATION	-0.208**	(0.105)	0.070	(0.064)		
LOG_TRANS	0.147***	(0.029)	-0.014	(0.019)		
LOG_IMP	-0.010	(0.019)	-0.018	(0.013)		

***, **, and * indicate a 99%, 95%, and 90% significance level.

This substitution effect has also been suggested by prior studies that examined the relationship between the organic results and sponsored ads (Jansen and Resnick 2006; Buscher et al. 2010; Jerath et al. 2014; Yin et al. 2014). Prior literature has shown that most consumers examine organic results first before sponsored ads (Jansen and Resnick 2006). An eye tracking study by Buscher et al. (2010) finds that, given the same keywords, when a list of less relevant ads is displayed, the organic results receive more visual attention, suggesting organic results may

substitute sponsored ads. Our result shows that sponsored ads may also substitute organic results. That is, when the organic results are diverse and thus less relevant because of keyword ambiguity, the sponsored ads may receive more attention. Evidently, in our robustness test presented in Online Appendix H, we find that higher keyword ambiguity is associated with less time spent on the organic listing, which supports our hypothesis of a substitution effect between organic results and sponsored ads.[14] It is important to note that we do not find evidence that consumers infer the sponsored ads as less relevant when they realize that the organic results are not very relevant, in the case of ambiguous keywords. Even if such a complementary effect might exist, it is dominated by the substitution effect between the two types of listings. Ghose and Yang (2009) also report that longer keywords, which may be less ambiguous, are associated with lower CTR. However, once keyword ambiguity is controlled for, we do not find statistically significant evidence for the impact of *NUM_WORDS* on the baseline CTR.

We present the coefficients of keyword characteristics on the decay in CTR with ad position in Column 2 of Table 4. First, the intercept term for $\beta_{1,k,t}$ is negative and statistically significant, indicating that CTR decreases with position. This finding is consistent with previous empirical studies that demonstrate strong position effects (e.g., Ghose and Yang 2009; Agarwal et al. 2011; Animesh et al. 2011). Second, we observe that keyword ambiguity (*TOPIC_ENTROPY*) has a statistically significant negative impact on the decay parameter $\beta_{1,k,t}$. That is, keywords with higher ambiguity seem to have lower $\beta_{1,k,t}$ and witness larger decreases in CTR with position. This finding indicates that although consumers are more likely to turn to sponsored ads when using more ambiguous keywords, they are less likely to continue clicking ads at lower positions. Since a consumer can evaluate whether an ad is relevant or not by reading the ad description (Jansen 2007), this reduction in click depth arises because there are fewer ads that are relevant to the consumer in the set of ads shown. When she realizes that fewer ads are

[14] Although we do not directly observe consumers' click behavior on organic search results, we are able to infer the time spent on the organic links by observing when a consumer starts a search session by entering a keyword, and when the consumer clicks on the first ad.

relevant, she gives up the search early, resulting in a faster decay in the CTR.[15] As the effects of keyword ambiguity on the baseline CTR and the decay parameter are opposite, we can conclude that our findings are not driven by consumer heterogeneity. Any effect driven by heterogeneity in consumer search costs would necessitate that keyword ambiguity had the same effect (either positive or negative) on both the baseline CTR and the decay parameter, which is inconsistent with our results.

Upon examining other keyword characteristics, we do not find statistically significant evidence for the effect of keyword length and presence of a brand name on the likelihood of a click. We find that the coefficient for *LOCATION* on $\beta_{0,k,t}$ is negative and statistically significant, indicating that keywords that have location information are less likely to get clicked than keywords that do not contain such information. The coefficient of *LOG_TRANS* is positive and statistically significant, indicating that transactional keywords are more likely to generate clicks. Additionally, we do not find a statistically significant impact of keyword popularity, as measured by *LOG_IMP*, on CTR. We also find that the number of ads displayed in an impression is positively correlated with CTR.

Estimates for Ad Position

Table 5 presents the results for ad position. The intercept term for $\phi_{1,k}$ is positive and statistically significant, indicating that $\overline{POS}_{a,-k_i}$ and POS_{ia} are highly positively correlated. More ads in the same impression seem to lead to a lower position, which is intuitive because, with more ads, the average position of all ads is lower. Some keyword characteristics such as *TOPIC_ENTROPY*, *LOG_TRANS*, and *LOG_IMP* also help explain the variation in ad position.

Table 6 shows the estimates for the covariance matrix Λ, which captures the unobserved correlation between CTR and position. The estimated variance (σ^2) is statistically significant, suggesting considerable variation in click performance across ads. The covariance estimate (σ_{12}) is statistically significant and positive, indicating a positive relationship between CTR

[15] We would like to thank the Senior Editor and the anonymous reviewers for this suggestion.

and ad position, which is consistent with previous literature (Agarwal et al. 2011; Ghose and Yang 2009).

Table 5: Estimation Results for Ad Position

	(1) Baseline ($\phi_{0,k}$)		(2) $\overline{POS}_{a,-k_i}$ ($\phi_{1,k}$)		(3) $NUM\_ADS(\phi_2)$	
Intercept	-1.854***	(0.045)	0.619***	(0.009)	0.436***	(0.006)
TOPIC_ENTROPY	0.208***	(0.061)	-0.083***	(0.021)		
NUM_WORDS	-0.001	(0.043)	0.018	(0.015)		
BRAND	0.058	(0.058)	-0.007	(0.020)		
LOCATION	0.091	(0.085)	-0.033	(0.029)		
LOG_TRANS	-0.096***	(0.024)	0.024***	(0.008)		
LOG_IMP	-0.016	(0.017)	0.018***	(0.006)		

***, **, and * indicate a 99%, 95%, and 90% significance level.

Summary of Main Results

Our analysis suggests that keyword ambiguity, as measured using topic entropy, has two opposing effects on the performance of ads associated with the keyword. Higher keyword ambiguity can lead to a higher baseline CTR; however, it also implies the CTR decreases more sharply with ad position. Taken together, the overall effect of keyword ambiguity on CTR for ads at various positions is a combination of these opposing effects. For example, the ads at positions 3-8 for a more ambiguous keyword tend to get a smaller number of clicks as compared to ads for a less ambiguous keyword.

Table 6: Estimation Results for Λ

Parameter	Estimate
σ_{12}	0.487*** (0.060)
σ^2	2.224*** (0.059)

***, **, and * indicate a 99%, 95%, and 90% significance level.

One of the strengths of our paper is the ability to perform cross-category analysis. Figure 5 illustrates how CTR changes with position and by topic. We use different colors to demonstrate different levels of CTR. Darker red colors represent higher CTR and light yellow colors represent lower CTR. The topics are ordered by CTR at the top position from the highest to the lowest. Our results suggest the position effect on CTR is heterogeneous across different topics.[16] For example, health-related keywords tend to attract higher CTR than car-related keywords at the top position, but the CTR decreases more quickly with positions.[17] The differences across keyword topics is an interesting phenomenon and understanding the factors that drive these differences is an important question that we leave for future research.

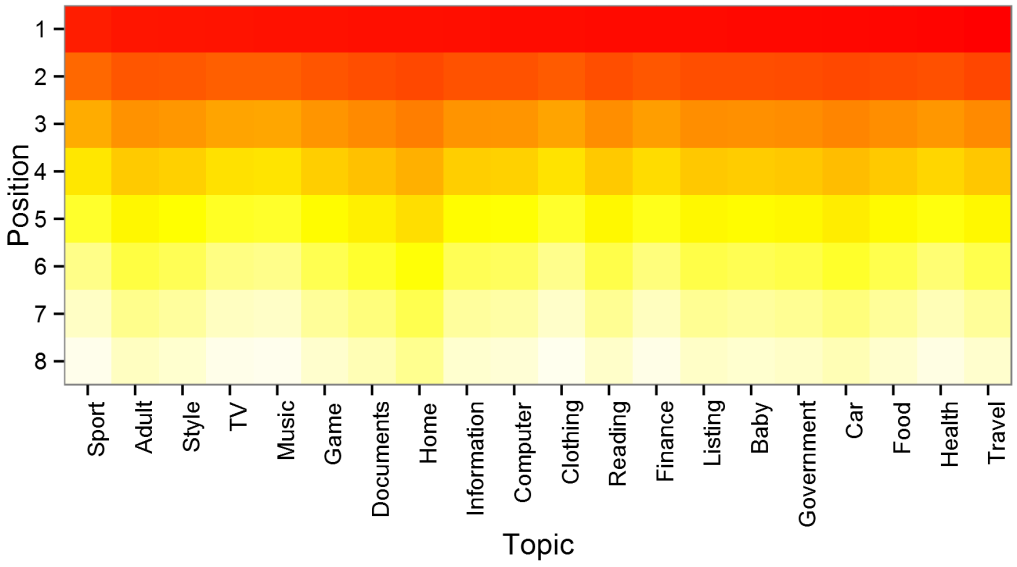

Figure 5: CTR by Topic for Sample Keywords

The large-scale, cross-category analysis presented here enables us to leverage data from multiple advertisers to generate insights that cannot be derived using data from a single

[16]Note that our model estimates the category level effects *after* controlling for the features of keywords that might be used for a certain category.

[17]To present the heterogeneous impact of position on CTR across topics, we also present box-plots of the topic specific intercepts (i.e., $\gamma_{0,t}$) and topic specific position effects (i.e., $\gamma_{1,t}$) in Online Appendix I.

advertiser. First, with a data set from a search engine that contains sponsored ads from different product categories and click-through activities with varied search interests, we are able to analyze how click performance varies by product category, and how the uncertainty of consumer search interests affects click performance. 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. It would not be possible to identify these effects separately if the data was provided by only one advertiser.

ROBUSTNESS CHECKS

Alternative Model Specifications

To check model performance, we compare our model against two alternative models. For convenience, we refer to Model 1 as our main model. Model 2 has a similar model specification as Model 1, but excludes the topic-specific intercepts (γ_t). 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 include topic entropy. We estimate the three alternative models using MCMC with our focal sample and compare the log-likelihood values based on the MCMC draws. Table 7 summarizes the differences among the three models. Comparing the average log-likelihood values, Model 1 outperforms the other two models (Table 7).

Prediction Performance

To compare the prediction performance of our model (Model 1) against the alternative models, we perform prediction based on both the estimation sample and a holdout sample. The holdout sample includes a different random sample of 10,000 impressions, which contains 1,493 unique keywords and 8,326 unique ads, resulting in 44,431 ads displayed. We predict both CTR and click-through (a binary outcome), and use mean absolute deviation (MAD) and root-mean-square error (RMSE) to evaluate prediction accuracy. Lower MAD and RMSE indicate better prediction accuracy. Table 8 reports the MAD and RMSE for each model based on CTR and click-through prediction. In all cases, our model (Model 1) outperforms the other

Table 7: Summary of Alternative Models

	Model 1	Model 2	Model 3
Description	Main model	No topic specific intercepts	No entropy or topic specific intercepts
Topic-specific intercept included	Yes	No	No
Entropy included	Yes	Yes	No
Control variables included	Yes	Yes	Yes
Average log-likelihood	**-3154.125**	-3175.537	-3182.824

two models, suggesting that incorporating topic heterogeneity and topic entropy provides better in-sample and out-of-sample prediction accuracy. The improvement is between 5.3%-9.3% on the training sample and 0.7%-6.7% on the holdout sample. We have also performed McNemar's test (Dietterich 1998) to compare Model 1 with Model 2, and Model 1 with Model 3, respectively. The results from McNemar's test suggest that the differences in accuracy between Model 1 and Model 2 and between Model 1 and Model 3 are statistically significant (p-value$<$0.01 for all pair-wise comparison tests).

Keywords with Zero Click

In our main analysis, we removed keywords that did not receive any click. To check how the removal of no-click keywords may affect the empirical results, we have randomly sampled 900 keywords that received no click, and compared the entropy values of these no-click keywords with keywords included in our sample. We found that the entropy values of the no-click keywords are slightly smaller than the keywords that received at least one click. A t-test also suggests that the difference in average entropy is statistically significant (p=0.0098). This is consistent with our main finding that more ambiguous keywords (with higher entropy) tend to have higher CTR.

Table 8: Model Comparison: Prediction Performance (Estimation and Holdout Samples)

		Model 1	Model 2	Model 3
Estimation Sample	MAD (CTR)	**0.039**	0.043	0.045
	RMSE (CTR)	**0.132**	0.144	0.148
	MAD (Click)	**0.039**	0.043	0.045
	RMSE (Click)	**0.197**	0.208	0.213
Holdout Sample	MAD (CTR)	**0.042**	0.045	0.045
	RMSE (CTR)	**0.147**	0.148	0.148
	MAD (Click)	**0.042**	0.045	0.045
	RMSE (Click)	**0.205**	0.212	0.213

Number of Google Organic Search Results

In the main analysis, we use the top-50-ranked Google organic search results to construct the corpus for topic modeling, because users rarely click on search results ranked below top 50 (Chitika 2013). Therefore, it is likely that Google will present the most relevant results on the first few pages, whereas the results that are listed in lower ranked pages can potentially be less relevant to users.

We check whether our empirical results are robust to the number of Google organic search results used for corpus construction by first re-estimating the topic model and re-computing topic entropy values based on different numbers of Google results (i.e., top-60, top-80, top-100). We then use the new values to re-estimate the hierarchical Bayesian model we have proposed. The results are presented in Online Appendix J. We find that both the entropy values (used to measure keyword ambiguity) and estimation results are robust to the number of organic search results used to construct the corpus for topic modeling.

DISCUSSION

Although practitioners have expressed concerns regarding bidding on ambiguous keywords in their search advertising campaigns because of the potential mismatch between advertisers'

intents and consumers' search interests,[18] no prior research has formally investigated the impact of keyword ambiguity on the performance of search advertising. This study is the first research that empirically examines how keyword ambiguity affects ad performance and how the effect varies across ad positions.

Managerial Implications

Our study has several managerial implications. Search advertising is one of the most dominant forms of online advertising, and has witnessed tremendous interest from advertisers. One of the biggest challenges faced by advertisers is keyword selection (Abhishek and Hosanagar 2007), i.e., which keywords should they incorporate into their advertising portfolio. Given the billions of unique keywords, this is an extremely challenging problem. Although researchers have proposed a few techniques such as multi-armed bandits (Rusmevichientong and Williamson 2006), it is difficult and expensive to search for and experiment with all potentially profitable keywords. Therefore, advertisers typically resort to heuristics to generate keywords.[19] The approach proposed in this paper to extract keyword semantic characteristics presents a new metric that can be used to evaluate keywords, along with other metrics as well as several insights that can be used to improve the selection of keywords. The CTR model proposed in the paper also improves the prediction of CTR by as much as 9.3% under certain cases, which can help advertisers improve the return on investment from these search advertising campaigns. As a typical advertiser uses several thousand keywords, this paper shows how unstructured big data techniques can be used by firms to improve their campaign performance by selecting the right keywords.[20] Our analysis also shows that keywords with a higher level of transactional intent tend to have higher CTR, which suggests that advertisers should bid on keywords with higher transactional intent to achieve higher click performance. Finally, since we can perform cross-category analysis, our paper is the first to offer insights into how CTR varies across

[18] http://www.positionresearch.com/keywords-a-strategic-approach/

[19] Although advertisers bid on several thousands keywords to learn about their performance, this is a very expensive strategy. They usually look for insights to generate relevant keywords that can reduce the experimentation costs.

[20] We thank the anonymous reviewer for suggesting that we make this link more explicit.

different industries and across different positions within a specific industry. For example, users are extremely likely to start clicking on travel related ads, but the drop off with position is faster. In contrast, for home related ads, although the baseline CTR is smaller, the drop off is much slower. It is worth noting that we only measure CTRs rather than conversion rates. To the degree that some firms may have a different objective, e.g., maximizing purchases or return on investment, it would be important for future research to examine the effect of these semantic characteristics of keywords on cost per click and conversion rate.[21]

This paper also provides implications for search engines. One of the most important reasons for the success of search advertising is providing advertisers the ability to target consumers based on their search interests through the use of keywords (Kiley 2006). However, keyword ambiguity may result in a mismatch between a consumer's search interest and an advertiser's intent, because the same keyword might be used by consumers with different search interests. Currently, search engines such as Google only provide advertisers an estimated CTR based on the keyword's prior performance and ad position, irrespective of how CTR may vary across topics.[22] The hierarchical Bayesian model we proposed in this paper suggests that, even without precise knowledge of each consumer's exact search interest, search engines can estimate topic-specific CTR as long as they have knowledge of the topic distribution of each keyword.[23] Therefore, they can develop a similar model to compute the predicted topic-specific CTR as a tool to aid advertisers when choosing potential keywords. As the advertisers know their own topic, they will be able to make a more informed decision while selecting keywords. In addition, search engines compute quality scores (QS) to weight the advertisers' bids and rank the sponsored ads which are a combination of historical CTR, relevance and landing page. However, our discussions with major search engines revealed that they have not considered

[21] We would like to thank an anonymous reviewer for this suggestion.

[22] https://support.google.com/adwords/answer/1659696?hl=en

[23] This is relatively easy for search engines to implement as they already have sophisticated algorithms to learn the topic distribution of each keyword. Because of the relatively low cost of Google API, advertisers can also easily collect Google organic search results for any potential keywords that they are interested in bidding on, and replicate the text mining techniques we have presented in this paper to estimate the topic distribution of each keyword.

scenarios where many different types of ads can be relevant (semantically) to a keyword (as in the case of ambiguous keywords) and QS alone might not be a good measure for such keywords. Incorporating keyword ambiguity in the ad serving decision might help the search engines improve revenues by showing different types of sponsored lists for ambiguous versus non-ambiguous keywords.

In Table 9, we further demonstrate how the hierarchical Bayesian model we develop can be used by search engines to predict CTR of a potential keyword. For each keyword in Table 9, we extract the semantic characteristics using text mining techniques. We then use the coefficients estimated from the hierarchical Bayesian model to predict the topic-specific baseline click propensity and decay parameter for each keyword. With the topic-specific baseline click propensity and decay parameter for each keyword, we can then predict the topic-specific CTR for each ad position. If search engines are interested in providing advertisers predicted CTR without prior knowledge of the topic of interest, they can use the topic probabilities extracted from the topic model as weights to obtain expected CTR for each position.[24] Therefore, the model we develop can be used by search engines to assist advertisers by predicting keyword performance both with and without specifying a particular topic that the advertiser might be interested in. These improvements for advertisers and search engines may increase the efficiency of search advertising and make it more attractive as it competes with newer forms of digital advertising such as mobile and video ads.

Finally, our results provide implications for search engines to provide better search results to users. First, our finding suggests a substitution effect between the organic listing and the sponsored listing. Currently, the two types of listings are managed by separate teams, which may work well if search engines' objective is to maximize the total number of clicks from sponsored ads. However, if search engines' objective is to increase user satisfaction or long-term retention, they can take into account the interplay between the two types of listings and optimize them

[24]That is, the predicted CTR of a keyword, unconditional on the topic, is a weighted average of topic-specific CTRs of the keyword. The weights are obtained from the topic model.

Table 9: Illustration of Predicted CTR by Topic for Sample Keywords

Keyword	Top 2 most related topic	Topic probability	Predicted CTR			
			Pos 1	Pos 2	Pos 3	Pos 4
car	Car	68.34%	15.00%	3.71%	0.93%	0.13%
	Finance	5.89%	14.25%	2.56%	0.40%	0.03%
	Overall (when the topic is unspecified)		14.67%	3.45%	0.82%	0.11%
car rental	Car	56.18%	15.00%	3.72%	1.15%	0.70%
	Travel	31.56%	17.07%	3.85%	0.99%	0.49%
	Overall (when the topic is unspecified)		15.51%	3.66%	1.05%	0.59%
car games	Game	63.81%	5.74%	1.70%	0.27%	0.10%
	Car	25.78%	7.01%	2.39%	0.46%	0.20%
	Overall (when the topic is unspecified)		6.12%	1.88%	0.32%	0.12%

jointly. Second, our finding suggests that more ambiguous keywords experience faster decay in CTR with ad position, as users find that the sponsored ads are less relevant compared to less ambiguous keywords. To alleviate the issue of ambiguity, search engines may prompt a list of categories (i.e., pre-generated topics based on the keyword) after users type in a keyword and ask users to choose a category of interest. By allowing users to associate a category with the keyword searched, search engines can then refine the listings to provide more relevant results.

Theoretical and Methodological Contributions

While the prior literature on search advertising largely focuses on a single advertiser or a few keywords from search engines, and implicitly assumes that there is no keyword ambiguity or mismatch between advertisers' intents and consumers' search interests, we relax this theoretical assumption and allow for potential keyword ambiguity in sponsored search. In particular, we achieve our goal by utilizing a data set from a major search engine that allows us to investigate the performance of keywords with varying levels of ambiguity. To our knowledge, this data set is one of the most extensive data sets used in the sponsored search literature that includes individual-level data across multiple product categories and advertisers.

The use of this data set, coupled with machine learning techniques and comprehensive empirical analyses, allows us to examine the effect of keyword ambiguity on keyword performance and make the following theoretical and methodological contributions. First, we contribute to the literature on search advertising by introducing a new keyword characteristic – topic entropy – that measures the ambiguity in the semantic meanings of a keyword. Although researchers have suggested the importance of understanding keyword ambiguity and ad relevance in analyzing the effectiveness of sponsored search advertising (e.g., Jansen and Resnick 2006; Jansen et al. 2007; Buscher et al. 2010; Athey and Ellison 2011; Jeziorski and Segal 2015), no prior study has operationalized the concept of keyword ambiguity and examined the impact of keyword ambiguity on search advertising performance. One challenge is to infer the different search interests associated with each keyword. To fill in this gap, we apply a machine learning based approach to measure keyword ambiguity. Specifically, we collect organic search results provided by search engines to better understand the semantic meanings of each keyword (Abhishek and Hosanagar 2007). This novel approach allows us to augment keyword meanings with easily available web data. We then apply a topic model on the organic search results to understand consumers' search interests when using different keywords, and subsequently quantify keyword ambiguity associated with each keyword by measuring the dispersion in keyword's topic distribution. Compared to previous studies that use human coders or a dictionary-based approach to measure the ambiguity of individual words, our approach is able to work with any word or phrase that consumers may use when searching online. Incidentally, the topic modeling approach allows us to obtain a distribution of topics for each keyword, which improves the prediction accuracy of the CTR model.

Second, we find that keyword ambiguity has a statistically significant impact on CTR. Specifically, we find that keyword ambiguity has a positive impact on top-positioned ads, but such effect diminishes with ad position. Our findings contribute to the search advertising literature towards a better understanding of the heterogeneity in the keyword performance through previously unobserved semantic dimensions.

Third, this paper increases our understanding of the interactions between organic results and sponsored ads by showing that for keywords with high ambiguity, consumers tend to substitute away from ambiguous organic results to sponsored ads, which leads to an increase in overall CTR for ambiguous keywords.

CONCLUSIONS

Using a unique data set from a major search engine that consists of detailed impression level click-through data across multiple categories, combined with the novel use of topic modeling to quantify keyword ambiguity, we analyze the relationship between keyword ambiguity and click-through performance using a hierarchical Bayesian approach that incorporates heterogeneity in CTR at keyword, topic, ad, and position levels, as well as potential endogeneity of ad position. We find that keyword ambiguity and keyword topics are significant predictors of keyword performance. Our results show that keywords with higher ambiguity are associated with higher CTR for top-positioned ads compared to those with lower ambiguity. In addition, we find that higher keyword ambiguity is associated with faster decay in CTR with ad position. Taken together, we find that the net effect of keyword ambiguity is a combination of these two opposing effects. In addition, we also observe that click-through performance varies significantly across topics. For example, topics such as sport receive fewer clicks, whereas categories such as travel receive more clicks.

There are a few limitations of our current work. One major limitation of our analysis is the lack of ad information. For example, we do not know ad characteristics such as ad copy, bid, and landing page. Although we try to control for unobserved ad-level differences in our hierarchical Bayesian model, 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 set. Therefore, our measure of keyword performance is limited only to CTR. From the perspective of search engines, CTR is a more relevant metric, because advertisers only pay when a click-through occurs. However, advertisers are not only interested in CTR, but also interested

in other measures of keyword performance such as conversion rate, return on investment as well as other non-transactional benefits such as increased awareness. A richer data set can be used to address this issue to examine the impact on other measures of keyword performance.

A third limitation is that we cannot track the same consumer over time. We thus make the assumption that each search impression is independent. Future research may incorporate the same consumer's potential dependency of search impressions in their model to improve prediction performance. Relatedly, although we have accounted for potential endogeneity of ad position using an instrumental variable approach and included ad heterogeneity (to control for ad level time-invariant factors such as ad popularity and profitability), topic heterogeneity (to control for topic specific popularity), and an extensive set of control variables, our identification of the effects of keyword ambiguity and ad position is limited by the innate limitations of observational data. It is possible that there may be other factors, such as consumer heterogeneity or supply-side factors (e.g., search engines' optimization) that are not accounted for in our model. The inability to track the same consumer over time and to randomize search results displayed to consumers may increase noise in our estimates. Although this is beyond the scope of our study, future research may employ behavioral laboratory experiments to tease out these dynamics.

Fourth, we measure keyword ambiguity based on Google organic search results that are observed at a time later than that of the click-through data examined in the empirical analysis. Insofar as organic and sponsored search results are typically managed by separate teams, it is unlikely that the search engines would optimize organic search results based on the performance of sponsored ads, which may alleviate the concern of reverse causality (i.e., the possibility that the performance of sponsored ads may affect keyword ambiguity, which is measured based on organic search results). It is worth noting that, while we have collected Google organic search results at different time points to show that the measure of keyword ambiguity is relatively stable, inevitably, there may be measurement errors that can introduce

biases into our empirical analysis. Ideally, access to data on organic and sponsored search results from the same impressions would help increase the accuracy of our findings.

Our study demonstrates the use of machine learning and computational linguistics to generate keyword characteristics such as keyword ambiguity, the presence of brand name and location, and the extent to which a keyword is transactional in a way that is fast, cheap, accurate, and meaningful. It also allows us to study the impact of these factors on consumer click behavior and provide implications for both researchers and practitioners to generate managerial insights with web content. We hope that the machine learning techniques and Bayesian analysis presented here provide a direction for both researchers and practitioners to explore the rich and meaningful web data towards a better understanding of search advertising. We believe that search advertising continues to evolve as an interesting area of information systems research, and this paper can contribute both methodologically and theoretically to this growing literature.

References

Abhishek, V., and Hosanagar, K. 2007. "Keyword Generation for Search Engine Advertising using Semantic Similarity between Terms," in *Proceedings of the Ninth International Conference on Electronic Commerce*.

Adelman, J. S., Brown, G. D., and Quesada, J. F. 2006. "Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times," *Psychological Science* (17:9), pp. 814–823.

Agarwal, A., Hosanagar, K., and Smith, M. D. 2011. "Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets," *Journal of Marketing Research* (48:6), pp. 1057–1073.

Agarwal, A., Hosanagar, K., and Smith, M. D. 2015. "Do Organic Results Help or Hurt Sponsored Search Performance?" *Information Systems Research* (26:4), pp. 695–713.

Agarwal, A., and Mukhopadhyay, T. 2016. "The Impact of Competing Ads on Click Performance in Sponsored Search," *Information Systems Research* (27:3), pp. 538–557.

Animesh, A., Ramachandran, V., and Viswanathan, S. 2010. "Research Note—Quality Uncertainty and the Performance of Online Sponsored Search Markets: An Empirical Investigation," *Information Systems Research* (21:1), pp. 190–201.

Animesh, A., Viswanathan, S., and Agarwal, R. 2011. "Competing "Creatively" in Sponsored Search Markets: The Effect of Rank, Differentiation Strategy, and Competition on Performance," *Information Systems Research* (22:1), pp. 153–169.

Aral, S., Ipeirotis, P., and Taylor, S. 2011. "Content and Context: Identifying the Impact of Qualitative Information on Consumer Choice," in *Proceedings of the 32th International Conference on Information Systems*, Shanghai, China: Association for Information Systems.

Arbatskaya, M. 2007. "Ordered Search," *The RAND Journal of Economics* (38:1), pp. 119–126.

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* (57:8), pp. 1485–1509.

Athey, S., and Ellison, G. 2011. "Position Auctions with Consumer Search," *The Quarterly Journal of Economics* (126:3), pp. 1213–1270.

Bao, Y., and Datta, A. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Management Science* (60:6), pp. 1371–1391.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993–1022.

Borowsky, R., and Masson, M. E. 1996. "Semantic Ambiguity Effects in Word Identification," *Journal of Experimental Psychology: Learning, Memory, and Cognition* (22:1), p. 63.

Broder, A. 2002. "A Taxonomy of Web Search," *Sigir Forum* (36:2), pp. 3–10.

Buscher, G., Dumais, S. T., and Cutrell, E. 2010. "The Good, the Bad, and the Random: An Eye-Tracking Study of Ad Quality in Web Search," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models," in *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada.

Che, H., Sudhir, K., and Seetharaman, P. 2007. "Bounded Rationality in Pricing under State-Dependent Demand: Do Firms Look Ahead, and If So, How Far?" *Journal of Marketing Research* (44:3), pp. 434–449.

Chen, Y., and He, C. 2011. "Paid Placement: Advertising and Search on the Internet," *The Economic Journal* (121:556), pp. F309–F328.

Chitika 2013. "The Value of Google Result Positionings," (available at https://chitika.com/google-positioning-value; accessed July 2017).

Dai, H. K., Zhao, L., Nie, Z., Wen, J.-R., Wang, L., and Li, Y. 2006. "Detecting Online Commercial Intention (OCI)," in *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, ACM.

Danescu-Niculescu-Mizil, C., Broder, A. Z., Gabrilovich, E., Josifovski, V., and Pang, B. 2010. "Competing for Users' Attention: on the Interplay between Organic and Sponsored Search Results," in *Proceedings of the 19th International Conference on World Wide Web*, ACM.

Dietterich, T. G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation* (10:7), pp. 1895–1923.

Edelman, B., Ostrovsky, M., and Schwarz, M. 2007. "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords," *American Economic Review* (97:1), pp. 242–259.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. 2007. "Duplicate Record Detection: A Survey," *IEEE Transactions on Knowledge and Data Engineering* (19:1), pp. 1–16.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2003. *Bayesian Data Analysis*, CRC press.

Ghose, A., and Ipeirotis, P. G. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering* (23:10), pp. 1498–1512.

Ghose, A., Ipeirotis, P. G., and Li, B. 2012. "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content," *Marketing Science*

(31:3), pp. 493–520.

Ghose, A., and Yang, S. 2009. "An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets," *Management Science* (55:10), pp. 1605–1622.

Goldenberg, J., Oestreicher-Singer, G., and Reichman, S. 2012. "The Quest for Content: How User-Generated Links Can Facilitate Online Exploration," *Journal of Marketing Research* (49:4), pp. 452–468.

Granka, L. A., Joachims, T., and Gay, G. 2004. "Eye-Tracking Analysis of User Behavior in WWW Search," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Griffiths, T. L., and Steyvers, M. 2004. "Finding Scientific Topics," *Proceedings of the National Academy of Science* (101:suppl 1), pp. 5228–5235.

Gu, B., Konana, P., Rajagopalan, B., and Chen, H.-W. M. 2007. "Competition Among Virtual Communities and User Valuation: The Case of Investing-Related Communities," *Information Systems Research* (18:1), pp. 68–85.

Hall, D., Jurafsky, D., and Manning, C. D. 2008. "Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics.

Hausman, J. A. 1996. "Valuation of New Goods under Perfect and Imperfect Competition," in *The Economics of New Goods*, University of Chicago Press, pp. 207–248.

Hoffman, P., Ralph, M. A. L., and Rogers, T. T. 2013. "Semantic Diversity: A Measure of Semantic Ambiguity Based on Variability in the Contextual Usage of Words," *Behavior Research Methods* (45:3), pp. 718–730.

Hoffman, P., and Woollams, A. M. 2015. "Opposing Effects of Semantic Diversity in Lexical and Semantic Relatedness Decisions," *Journal of Experimental Psychology: Human Perception and Performance* (41:2), pp. 385–402.

Interactive Advertising Bureau 2017. "IAB Internet Advertising Revenue Report, 2016 Full Year Results," (available at http://www.iab.com/insights/iab-internet-advertising-revenue-

report-conducted-by-pricewaterhousecoopers-pwc-2/; accessed July 2017).

Jansen, B. J. 2007. "The Comparative Effectiveness of Sponsored and Nonsponsored Links for Web e-Commerce Queries," *ACM Transactions on the Web* (1:1), p. 3.

Jansen, B. J., Brown, A., and Resnick, M. 2007. "Factors Relating to the Decision to Click on a Sponsored Link," *Decision Support Systems* (44:1), pp. 46–59.

Jansen, B. J., and Resnick, M. 2006. "An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links during Ecommerce Web Searching," *Journal of the American Society for Information Science and Technology* (57:14), pp. 1949–1961.

Jastrzembski, J. E. 1981. "Multiple Meaning, Number of Related Meanings, Frequency of Occurrence, and the Lexicon," *Cognitive Psychology* (13:2), pp. 278–305.

Jerath, K., Ma, L., and Park, Y.-H. 2014. "Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity," *Journal of Marketing Research* (51:4), pp. 480–486.

Jeziorski, P., and Segal, I. 2015. "What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising," *American Economic Journal: Microeconomics* (7:3), pp. 24–53.

Kai, L. 1998. "Bayesian Inference in a Simultaneous Equation Model with Limited Dependent Variables," *Journal of Econometrics* (85:2), pp. 387–400.

Kellas, G., Ferraro, F. R., and Simpson, G. B. 1988. "Lexical Ambiguity and the Timecourse of Attentional Allocation in Word Recognition," *Journal of Experimental Psychology: Human Perception and Performance* (14:4), p. 601.

Kiley, D. 2006. "Google's Search for the Advertising Edge," (available at http://www.bloomberg.com/news/articles/2006-01-18/googles-search-for-the-advertising-edge; accessed May 2017).

Landauer, T. K., and Dumais, S. T. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge,"

Psychological Review (104:2), p. 211.

Lee, D., Hosanagar, K., and Nair, H. 2016. "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," Working paper.

Levenshtein, V. I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady* (10), pp. 707–710.

McCallum, A. K. 2002. "MALLET: A Machine Learning for Language Toolkit," Http://mallet.cs.umass.edu.

McDonald, S. A., and Shillcock, R. C. 2001. "Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing," *Language and Speech* (44:3), pp. 295–322.

Moe, W. W. 2003. "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology* (13:1).

Moorthy, S., Ratchford, B. T., and Talukdar, D. 1997. "Consumer Information Search Revisited: Theory and Empirical Analysis," *Journal of Consumer Research* (23:4), pp. 263–277.

Narayanan, S., and Kalyanam, K. 2015. "Position Effects in Search Advertising and Their Moderators: A Regression Discontinuity Approach," *Marketing Science* (34:3), pp. 388–407.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. 2012. "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science* (31:3), pp. 521–543.

Nevo, A. 2000. "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *RAND Journal of Economics* (31:3), pp. 395–421.

Nevo, A. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica* (69:2), pp. 307–342.

Rayner, K., and Duffy, S. A. 1986. "Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity," *Memory & Cognition* (14:3), pp. 191–201.

Rodd, J., Gaskell, G., and Marslen-Wilson, W. 2002. "Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access," *Journal of Memory and Language* (46:2), pp.

245–266.

Rodd, J. M., Gaskell, M. G., and Marslen-Wilson, W. D. 2004. "Modelling the Effects of Semantic Ambiguity in Word Recognition," *Cognitive Science* (28:1), pp. 89–104.

Rusmevichientong, P., and Williamson, D. P. 2006. "An Adaptive Algorithm for Selecting Profitable Keywords for Search-based Advertising Services," in *Proceedings of the 7th ACM Conference on Electronic Commerce*.

Rutz, O. J., and Bucklin, R. E. 2011. "From Generic to Branded: A Model of Spillover in Paid Search Advertising," *Journal of Marketing Research* (48:1), pp. 87–102.

Rutz, O. J., Bucklin, R. E., and Sonnier, G. P. 2012. "A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising," *Journal of Marketing Research* (49:3), pp. 306–319.

Rutz, O. J., and Trusov, M. 2011. "Zooming In on Paid Search Ads–A Consumer-Level Model Calibrated on Aggregated Data," *Marketing Science* (30:5), pp. 789–800.

Rutz, O. J., Trusov, M., and Bucklin, R. E. 2011. "Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?" *Marketing Science* (30:4), pp. 646–665.

Saad, G., and Russo, J. E. 1996. "Stopping Criteria in Sequential Choice," *Organizational Behavior and Human Decision Processes* (67:3), pp. 258–270.

Singh, P. V., Sahoo, N., and Mukhopadhyay, T. 2014. "How to Attract and Retain Readers in Enterprise Blogging ?" *Information Systems Research* (25:1), pp. 35–52.

Steyvers, M., and Griffiths, T. 2007. "Probabilistic Topic Models," *Handbook of Latent Semantic Analysis* (427:7), pp. 424–440.

Szymanski, B. K., and Lee, J.-S. 2006. "Impact of ROI on Bidding and Revenue in Sponsored Search Advertisement Auctions," in *Second Workshop on Sponsored Search Auctions*.

Varian, H. R. 2007. "Position Auctions," *International Journal of Industrial Organization* (25:6), pp. 1163–1178.

Weber, T. A., and Zheng, Z. E. 2007. "A Model of Search Intermediaries and Paid Referrals."

Information Systems Research (18:4), pp. 414–436.

Weitzman, M. L. 1979. "Optimal Search for the Best Alternative," *Econometrica* (47:3), pp. 641–54.

Yang, S., and Ghose, A. 2010. "Analyzing the Relationship Between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?" *Marketing Science* (29:4), pp. 602–623.

Yao, S., and Mela, C. F. 2011. "A Dynamic Model of Sponsored Search Advertising," *Marketing Science* (30:3), pp. 447–468.

Yin, D., Mei, S., Cao, B., Sun, J.-T., and Davison, B. D. 2014. "Exploiting Contextual Factors for Click Modeling in Sponsored Search," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*.

ONLINE APPENDIX A
SUMMARY OF EMPIRICAL STUDIES ON SPONSORED SEARCH ADVERTISING

Paper	**Goal**	**Data Source**	**Industry**	**Level of detail**	**Number of keywords examined**
Agarwal et al. (2011)	Impact of position on click-through and conversion	Advertiser	Pet products	Aggregate	68
Agarwal et al. (2015)	Impact of organic competition on click-through and conversion	Advertiser	Pet products	Aggregate	36
Chan et al. (2011)	Measuring the value of customers acquired from sponsored search	Advertiser	Lab supplies	Individual	90-208
Chan and Park (2015)	Advertiser valuation of consumer search activities	Search engine	Sporting goods	Individual	1
Ghose and Yang (2009)	Impact of keyword attributes on click-through and conversion	Advertiser	Retail	Aggregate	1,878
Goldfarb and Tucker (2011)	Online and offline advertising channel substitution	Advertiser	Legal service	Aggregate	139
Jerath et al. (2014)	Impact of keyword popularity on click performance	Search engine		Individual	1,200

Jeziorski and Segal (2015)	Quantifying rational user experience and externalities among ads	Search engine		Individual	4
Rutz and Bucklin (2011)	Spill-over from generic to branded keywords	Advertiser	Hotel	Aggregate	Several hundred
Rutz and Trusov (2011)	Effects of ad attributes on ad performance	Advertiser	Ringtone	Aggregate	80
Rutz et al. (2011)	Quantifying indirect effects of paid search	Advertiser	Automotive	Aggregate	3,186
Rutz et al. (2012)	Impact of ad position on conversion performance	Advertiser	Hotel	Aggregate	301
Yang and Ghose (2010)	Relationship between organic and sponsored search	Advertiser	Retail	Aggregate	426
Yang et al. (2014)	Impact of ad competition on click performance and cost per click	Advertiser	Digital camera and video products	Aggregate	1,573
Yao and Mela (2011)	Modeling user, advertiser, and search engine interaction	Search engine	Music management	Individual	

ONLINE APPENDIX B

LATENT DIRICHLET ALLOCATION

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. 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 our discussion on corpus construction, 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 ϕ_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 θ_{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 $w_{dn} \sim Multinomial(\phi_{t_{dn}})$.

ψ and ω are hyper-parameters for the two prior distributions - $Dirichlet(\psi)$ as the prior distribution of ϕ (word distribution in a topic) and $Dirichlet(\omega)$ as the prior distribution of θ (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 ϕ and θ. 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.

ONLINE APPENDIX C

TOPIC DISTRIBUTION OF SAMPLE KEYWORDS

Figure A.1 illustrates the topic distribution of some sample keywords. In Figure A.1, topics are labeled on the horizontal axis, and keywords are labeled on the vertical axis. The size of 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.

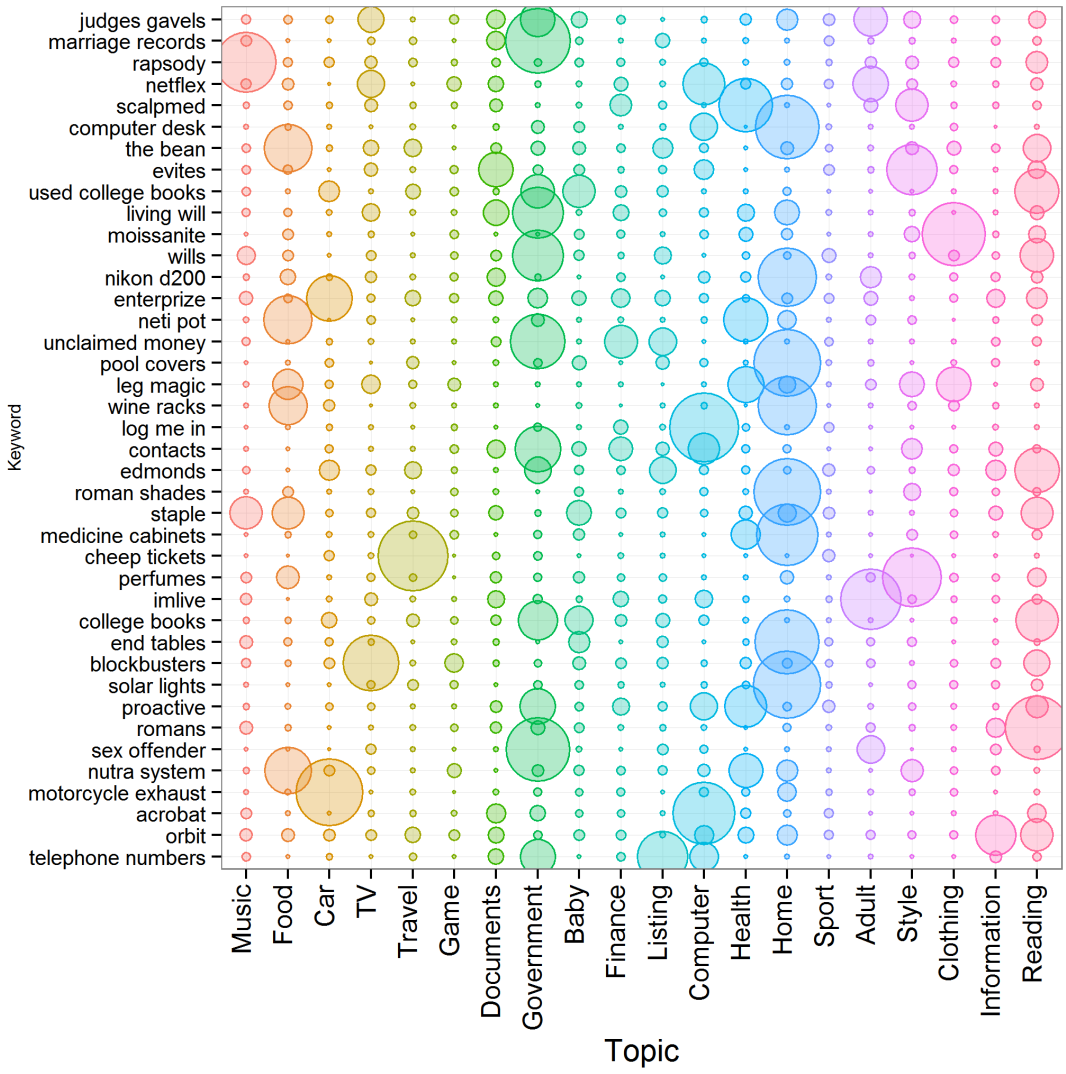

Figure A.1: Topic Distribution of Sample Keywords

ONLINE APPENDIX D

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.[25] 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 (i.e., Levenshtein distance computed from our model) is greater than 0.85; (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 "chipo**lt**e" with the brand "chipo**tl**e," and "walmart" with "wal mart." [26]

We use a similar approach to extract whether a keyword contains location information. We obtain a list of U.S. 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 partial 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.[27]

[25] 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.

[26] 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.

[27] 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.

ONLINE APPENDIX E

EXTRACTING TRANSACTIONAL INTENT

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. 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 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. These transactional words are listed in Table A.1. Then, for each search 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 A.1: Transactional Words

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

ONLINE APPENDIX F

CORRELATION AMONG VARIABLES

Table A.2: Correlation among Variables

Variable	*TOPIC_ENTROPY*	*NUM_WORDS*	*BRAND*	*LOCATION*	*LOG_TRANS*	*LOG_IMP*
TOPIC_ENTROPY	1.00					
NUM_WORDS	-0.38	1.00				
BRAND	-0.03	0.06	1.00			
LOCATION	0.00	0.16	0.04	1.00		
LOG_TRANS	0.07	-0.04	0.22	0.01	1.00	
LOG_IMP	-0.03	-0.17	0.19	-0.02	0.08	1.00

ONLINE APPENDIX G

THE GIBBS SAMPLING PROCEDURE

We estimate our hierarchical Bayesian model using a Gibbs sampling procedure, which samples parameters iteratively from their conditional distributions given the data and all other parameters. Note that we model CTR conditional on the topic t_i related to impression i. As t_i is not observed, we use a data augmentation approach by simulating topic assignment based on membership probabilities $\hat{\theta}_{k_i} = (\hat{\theta}_{k_i,1}, ..., \hat{\theta}_{k_i,T})'$, which is estimated from a topic model. In addition, U_{iat} is also a latent variable that involves data augmentation.

For simplification, we assume that

$\beta^{kt} = (\beta_{0,k,t}, \beta_{1,k,t})'$,

$\phi^k = (\phi_{0,k}, \phi_{1,k})'$,

$u_k^\beta = (u_{0,k}, u_{1,k})'$,

$\Delta^\beta = (\Delta_0^\beta, \Delta_1^\beta)'$,and

$\chi^k = \Delta^\beta X_k + u_k^\beta$.

The hierarchical Bayesian model can be written in the following hierarchical form:$t_i|\hat{\theta}_{k_i}$

$U_{iat}|POS_{ia}; \beta^{k_i,t}, \beta_2, \tau_a, \eta_{ia}$

$POS_{ia}|\phi_{0,k_i}, \phi_{1,k_i}, \phi_2, \varepsilon_{ia}$

$\eta_{ia}, \varepsilon_{ia}|\Lambda$

$\tau_a|\nu^\tau$

$\beta^{k_i,t}|\gamma_t, \Delta^\beta, \Omega^\beta$

$\phi^{k_i}|\Delta^\phi, \Omega^\phi$

$\gamma_t|\Psi$

We assume the following prior specifications:

$\beta_2 \sim N(0, v^{\beta_2})$

$\phi_2 \sim N(0, v^{\phi_2})$

$vecr(\Delta^{\beta}) \sim MVN(\overline{\Delta^{\beta}}, A^{\beta})$

$vecr(\Delta^{\phi}) \sim MVN(\overline{\Delta^{\phi}}, A^{\phi})$

$v^{\tau} \sim IG(m, n)$

$\sigma_{12} \sim N(r_0, b_0)\sigma^2 \sim IG(\frac{v_0}{2}, \frac{c_0}{2})$

$\Psi \sim IW(v^{\Psi}, V^{\Psi})$

$\Omega^{\beta} \sim IW(v^{\beta}, V^{\beta})$

$\Omega^{\phi} \sim IW(v^{\phi}, V^{\phi})$

We describe the Gibbs sampling procedure below.

Step 1. Draw $t_i \sim Multinomial(\hat{\theta}_{k_i})$ ***for each impression*** i***.***

Step 2. Draw U_{iat} ***for each observation.***

We can draw U_{iat} from the following posterior distribution:

$U_{iat} \sim TN(\mu_{iat}, \sigma_{1|2})$,

where TN denotes the truncated normal distribution, and U_{iat} is truncated above zero if $Click_{ia} = 1$, and below zero if 0. Let $\overline{U}_{iat} = \beta_{0,k_i,t} + \beta_{1,k_i,t} POS_{ia} + \beta_2 NUM\_AD_i + \tau_a$, $\tilde{\varepsilon}_{ia} = POS_{ia} - \phi_{0,k_i} - \phi_{1,k_i}\overline{POS}_{a,-k_i} - \phi_2 NUM\_AD_i$, then

$\mu_{iat} = \overline{U}_{iat} + \frac{\sigma_{12}}{\sigma^2+\sigma_{12}}\tilde{\varepsilon}_{ia}$,

$\sigma_{1|2} = 1 - \frac{\sigma_{12}}{\sigma^2+\sigma_{12}} = \frac{\sigma^2}{\sigma^2+\sigma_{12}}$.

Step 3. Draw χ^k, ϕ^k for each keyword k.

For each keyword k, let N_k be the number of observations such that $k_i = k$. Let

$\Gamma_k = (\chi^k, \phi^k)'$,

$z_{1ia} = (1, POS_{ia})'$,

$z_{2ia} = (1, \overline{POS}_{a,-k_i})'$,

$y_{1ia} = U_{iat} - z'_{1ia}\gamma_t - \beta_2 NUM\_AD_i - \tau_a$,

$y_{2ia} = POS_{ia} - \phi_2 NUM\_AD_i$.

Then

$y_{1ia} = z'_{1ia}\chi^k + \eta_{ia}$,

$y_{2ia} = z'_{2ia}\phi^k + \varepsilon_{ia}$,

where $(\eta_{ia}, \varepsilon_{ia})' \sim MVN(0, \Lambda)$. We can write it in matrix version as

$y_{1k} = Z'_{1k}\chi^k + \eta_k$,

$y_{2k} = Z'_{2k}\phi^k + \varepsilon_k$,

or more compactly, as

$Y = Z'_k\Gamma_k + E_k$,

where $Y_k = (y_{1k}, y_{2k})'$, $Z_k = \begin{pmatrix} Z_{1k} & 0 \\ 0 & Z_{2k} \end{pmatrix}$, and $E_k = (\eta_k, \varepsilon_k)' \sim MVN(0, \Lambda \otimes I_{N_k})$.

We can rewrite $\Delta^\beta = \begin{pmatrix} \Delta^\beta_{11} & \cdots & \Delta^\beta_{1r} \\ \Delta^\beta_{21} & \cdots & \Delta^\beta_{2r} \end{pmatrix}$ as a vector $\delta^\beta = vecr(\Delta^\beta) = (\Delta^\beta_{11}, \ldots, \Delta^\beta_{1r,} \Delta^\beta_{21}, \ldots, \Delta^\beta_{2r})'$.
Similarly, $\delta^\phi = vecr(\Delta^\phi)$.

With prior distribution $\Gamma_k \sim MVN(\overline{\Gamma}_k, \psi_0)$, where $\overline{\Gamma}_k = [I_4 \otimes X_k'] \begin{pmatrix} \delta^\beta \\ \delta^\phi \end{pmatrix}$, and

$\psi_0 = \begin{pmatrix} \Omega^\beta & 0 \\ 0 & \Omega^\phi \end{pmatrix}$, we can draw Γ_k from the following posterior distribution:

$\Gamma_k | all\ other\ parameters \sim MVN(\widetilde{\Gamma}_k, \widetilde{\psi})$,

where $\widetilde{\Gamma}_k = \widetilde{\psi}[Z_k'(\Lambda^{-1} \otimes I_{N_k})Y_k + \psi_0^{-1}\overline{\Gamma}_k]$, and $\widetilde{\psi} = [Z_k'(\Lambda^{-1} \otimes I_{N_k})Z_k + \psi_0^{-1}]^{-1}$.

Step 4. Draw γ_t for each topic t.

For each topic t, let N_t be the number of observations with $t_i = t$. Let $Z_{ia} = (1, POS_{ia})'$, $\tilde{\varepsilon}_{ia} = POS_{ia} - \phi_{0,k_i} - \phi_{1,k_i} - \phi_2 NUM\_AD_i$, $U^1_{iat} = U_{iat} - Z_{ia}\chi^k - \beta_2 NUM\_AD_i - \tau_a - \frac{\sigma_{12}}{\sigma^2 + \sigma^2_{12}}\tilde{\varepsilon}_{ia}$, and $\sigma_{1|2} = \frac{\sigma^2}{\sigma^2 + \sigma^2_{12}}$. Then $U^1_{iat} \sim N(Z_{ia}'\gamma_t, \sigma_{1|2})$. We can write it in matrix version as

$U^{1t} \sim MVN(Z^t\gamma_t, \sigma_{1|2} I_{N_t})$,

where $U^{1t} : N_t \times 1$ includes all U^1_{iat} such that $t_i = t$, and Z^t is a $N_t \times 2$ matrix. With prior $\gamma_t \sim MVN(0, \Psi)$, we then draw γ_t from the following posterior distribution:

$\gamma_t | all\ other\ parameters \sim MVN(\widetilde{\gamma_t}, \widetilde{\Psi})$,

where $\widetilde{\Psi} = [(\Psi)^{-1} + (Z^t)'Z^t/\sigma_{1|2}]^{-1}$, and $\widetilde{\gamma_t} = \widetilde{\Psi}[(Z^t)'U^{1t}/\sigma_{1|2}]$.

Step 5. Draw τ_a for each ad a.

For each ad a, let n_a be the number of observations. We define

$\tilde{\varepsilon}_{ia} = POS_{ia} - \phi_{0,k_i} - \phi_{1,k_i} - \phi_2 NUM\_AD_i$,

$U^2_{iat} = U_{iat} - (\beta_{0,k_i,t} + \beta_{1,k_i,t} POS_{ia}) - \beta_2 NUM\_AD_i - \frac{\sigma_{12}}{\sigma^2+\sigma^2_{12}}\tilde{\varepsilon}_{ia}$, and $\sigma_{1|2} = \frac{\sigma^2}{\sigma^2+\sigma^2_{12}}$. Then

$U^2_{iat} \sim N(\tau_a, \sigma_{1|2})$. With prior $\tau_a \sim N(0, v^\tau)$, the posterior distribution of τ_a is

$\tau_a | all\ other\ parameters \sim MVN(\widetilde{\tau_a}, \widetilde{v^\tau})$,

where $\widetilde{\tau_a} = \frac{n_a v^\tau \overline{U^2_{iat}}}{n_a v^\tau + \sigma_{1|2}}$, and $\widetilde{v^\tau} = \frac{\sigma_{1|2} v^\tau}{n_a v^\tau + \sigma_{1|2}}$.

Step 6. Draw β_2.

Let $\tilde{\varepsilon}_{ia} = POS_{ia} - \phi_{0,k_i} - \phi_{1,k_i} - \phi_2 NUM\_AD_i$,

$U^3_{iat} = U_{iat} - (\beta_{0,k_i,t} + \beta_{1,k_i,t} POS_{ia}) - \beta_2 NUM\_AD_i - \frac{\sigma_{12}}{\sigma^2+\sigma^2_{12}}\tilde{\varepsilon}_{ia}$, and $X_i = NUM\_AD_i$, then

$U^3_{iat} \sim N(\beta_2 X_i, \sigma_{1|2})$. With prior $\beta_2 \sim N(0, v^{\beta_2})$, the posterior distribution of β_2 is

$\beta_2 | all\ other\ parameters \sim N(\widetilde{\beta_2}, \widetilde{v^{\beta_2}})$,

where $\widetilde{v^{\beta_2}} = [\sigma^{-1}_{1|2} X'X + (v^{\beta_2})^{-1}]^{-1}$, and $\widetilde{\beta_2} = v^{\beta_2}[\sigma^{-1}_{1|2} X'U^3]$.

Step 7. Draw ϕ_2.

Let $\tilde{\eta}_{ia} = U_{iat} - (\beta_{0,k_i,t} + \beta_{1,k_i,t} POS_{ia}) - \beta_2 NUM\_AD_i - \tau_a$,

$w_{ia} = POS_{ia} - (\phi_{0,k_i} + \phi_{1,k_i}\overline{POS}_{a,-k_i}) - \sigma_{12}\tilde{\eta}_{ia}$, $X_i = NUM\_AD_i$, then $w_{ia} \sim N(\phi_2 X_i, \sigma^2)$. With prior $\phi_2 \sim N(0, v^{\phi_2})$, the posterior distribution of ϕ_2 is

$\phi_2 | all\ other\ parameters \sim N(\widetilde{\phi_2}, \widetilde{v^{\phi 2}})$,

where $\widetilde{v^{\phi_2}} = [\sigma^{-2}X'X + (v^{\phi_2})^{-1}]^{-1}$, and $\widetilde{\phi_2} = v^{\phi_2}[\sigma^{-2}X'w]$.

Step 8. Draw Δ^β.

Let K be the number of keywords. With $\Delta^\beta = \begin{pmatrix} \Delta^\beta_{11} & \cdots & \Delta^\beta_{1r} \\ \\ \Delta^\beta_{21} & \cdots & \Delta^\beta_{2r} \end{pmatrix}$, we have

$\delta^\beta = vecr(\Delta^\beta) = (\Delta^\beta_{11},\ldots,\Delta^\beta_{1r},\Delta^\beta_{21},\ldots,\Delta^\beta_{2r})'$. Therefore, $\chi^k = \begin{pmatrix} X'_k & 0 \\ \\ 0 & X'_k \end{pmatrix} \delta^\beta + u^\beta_k$,where

$u^\beta_k \sim MVN(0,\Omega^\beta)$. We can rewrite this in matrix format as

$$\begin{pmatrix} \chi_0 \\ \\ \chi_1 \end{pmatrix} = \begin{pmatrix} X & 0 \\ \\ 0 & X \end{pmatrix} \delta^\beta + E,$$

where $\chi_0 = (\chi_{01},\ldots,\chi_{0K})'$, $\chi_1 = (\chi_{11},\ldots,\chi_{1K})'$, $X = (X_1,\ldots,X_K)'$, and $E \sim MVN(0,\Omega^\beta \otimes I_K)$. More compactly, we can write

$$\chi = (I_2 \otimes X)\delta^\beta + E.$$

With prior $\delta^\beta \sim MVN(\overline{\Delta^\beta},A^\beta)$, we can draw δ^β from the following posterior distribution:

$$\delta^\beta | all\ other\ parameters \sim MVN(\widetilde{\Delta^\beta},\widetilde{A^\beta}),$$

where $\widetilde{A^\beta} = [(I_2 \otimes X)'((\Omega^\beta)^{-1} \otimes I_K)(I_2 \otimes X) + (A^\beta)^{-1}]^{-1}$, and
$\widetilde{\Delta^\beta} = \widetilde{A^\beta}[(I_2 \otimes X)'((\Omega^\beta)^{-1} \otimes I_K)\chi + (A^\beta)^{-1}\overline{\Delta^\beta}]$.

Step 9. Draw Δ^β.

Similar to the previous step, with prior $\delta^\phi \sim MVN(\overline{\Delta^\phi},A^\phi)$, we can draw δ^ϕ from the following posterior distribution:

$\delta^{\phi}|all\ other\ parameters \sim MVN(\widetilde{\Delta^{\phi}}, \widetilde{A^{\phi}})$,

where $\widetilde{A^{\phi}} = [(I_2 \otimes X)'((\Omega^{\phi})^{-1} \otimes I_K)(I_2 \otimes X) + (A^{\phi})^{-1}]^{-1}$, and
$\widetilde{\Delta^{\phi}} = \widetilde{A^{\phi}}[(I_2 \otimes X)'((\Omega^{\phi})^{-1} \otimes I_K)\chi + (A^{\phi})^{-1}\overline{\Delta^{\phi}}]$.

Step 10. Draw v^{τ}.

With prior $v^{\tau} \sim IG(m,n)$, we can draw v^{τ} from its posterior distribution $IG(m+\frac{A}{2}, n+\frac{\tau'\tau}{2})$, where A is the total number of unique ads.

Step 11. Draw σ_{12}.

With prior $\sigma_{12} \sim N(0, b_0)$, we can draw σ_{12} from its posterior distribution $N(\widetilde{r}, \tilde{b})$, where $\tilde{r} = (\sigma^{-2}\tilde{\eta}'\tilde{\varepsilon})$, and $\tilde{b} = (\sigma^{-2}\tilde{\eta}'\tilde{\eta} + b_0^{-1})^{-1}$.

Step 12. Draw σ^2.

With prior $\sigma^2 \sim IG(v_0, c_0)$, we can draw σ^2 from its posterior distribution $IG(\tilde{v}, \tilde{c})$, where $\tilde{v} = v_0 + \frac{N}{2}$, and $\tilde{c} = c_0 + \frac{(\tilde{\varepsilon}-\tilde{\eta}\sigma_{12})'(\tilde{\varepsilon}-\tilde{\eta}\sigma_{12})}{2}$. N is the number of observations.

Step 13. Draw Ω^{β}.

With prior $\Omega^{\beta} \sim IW(v^{\beta}, V^{\beta})$, we can draw Ω^{β} from its posterior distribution:

$IW(v^{\beta} + K, V^{\beta} + \sum_{k=1}^{K}(\chi^k - \Delta^{\beta}X_k)(\chi^k - \Delta^{\beta}X_k)')$,

where K is the number of keywords.

Step 14. Draw Ω^{ϕ}.

Similarly, with prior $\Omega^{\phi} \sim IW(v^{\phi}, V^{\phi})$, we can draw Ω^{ϕ} from its posterior distribution:

$IW(v^{\phi} + K, V^{\phi} + \sum_{k=1}^{K}(\phi^k - \Delta^{\phi}X_k)(\phi^k - \Delta^{\phi}X_k)')$.

Step 14. Draw Ψ.

With prior $\Psi \sim IW(v^{\Psi}, V^{\Psi})$, we can draw Ψ from its posterior distribution:

$$IW(v^{\Psi} + T, V^{\Psi} + \sum_{t=1}^{T} \gamma_t \gamma_t'),$$

where T is the number of topics.

ONLINE APPENDIX H

ANALYSIS ON TIME BEFORE FIRST CLICK

Although we do not directly observe consumers' click behavior on organic search results, we are able to infer the time on the organic links by observing when a consumer starts a search session by entering a keyword, and when the consumer clicks on the first ad. Therefore, we focus our analysis on the subset of consumers who have clicked on at least one sponsored ad. We denote *DURATION* as the time between the consumer starts a search session and makes the first click, and we use *DURATION* as a proxy for the time spent on the organic listing. We run a linear regression of *DURATION* on the keyword attributes of interests. As shown in Table A.3, higher keyword ambiguity is associated with less time spent on organic search results, while more precise keywords tend to attract more attention on organic search results. This result provides partial evidence that a more ambiguous keyword may reduce the attractiveness of organic search results, and consumers may turn to sponsored ads for finding an alternative that meets their needs.

Table A.3: Regression Results: Time before First Click

Variable	Estimates	
Intercept	56.925***	(0.675)
TOPIC_ENTROPY	-4.643***	(0.232)
NUM_WORDS	0.335**	(0.155)
BRAND	-2.068***	(0.210)
LOCATION	-1.200***	(0.301)
LOG_TRANS	-2.895***	(0.090)
LOG_IMP	-2.674***	(0.057)
Observations	551,239	

***, **, and * indicate a 99%, 95%, and 90% significance level.

ONLINE APPENDIX I

BOXPLOTS OF TOPIC SPECIFIC EFFECTS

To present the heterogeneous impact on CTR across topics, we overlay boxplots of the topic specific intercepts (i.e., $\gamma_{0,t}$) in Figure A.2 and the topic specific effects of position (i.e., $\gamma_{1,t}$) in Figure A.3. Because we have mean-centered all keyword characteristics when estimating the hierarchical Bayesian model, $\gamma_{0,t}$ and $\gamma_{1,t}$ can be interpreted as estimates of $\beta_{0,k,t}$ and $\beta_{1,k,t}$ for a typical keyword in topic t of which the covariates are set to mean values. As we can see from Figure A.2, the means of the posterior distribution of the topic specific intercepts are highest for topics "travel" and "health," suggesting that consumers who are interested in those topics may be more likely to click on ads at top positions. In contrast, for topics "sport" and "adult," CTR is lower at top positions.

As we can see from Figure A.3, the means of the posterior distribution of the topic specific effects of position are highest for topics "home" and "documents," suggesting that consumers who are interested in those topics may be more likely to click on ads at lower positions. In contrast, for topics "music" and "clothing," CTR decreases faster with position.

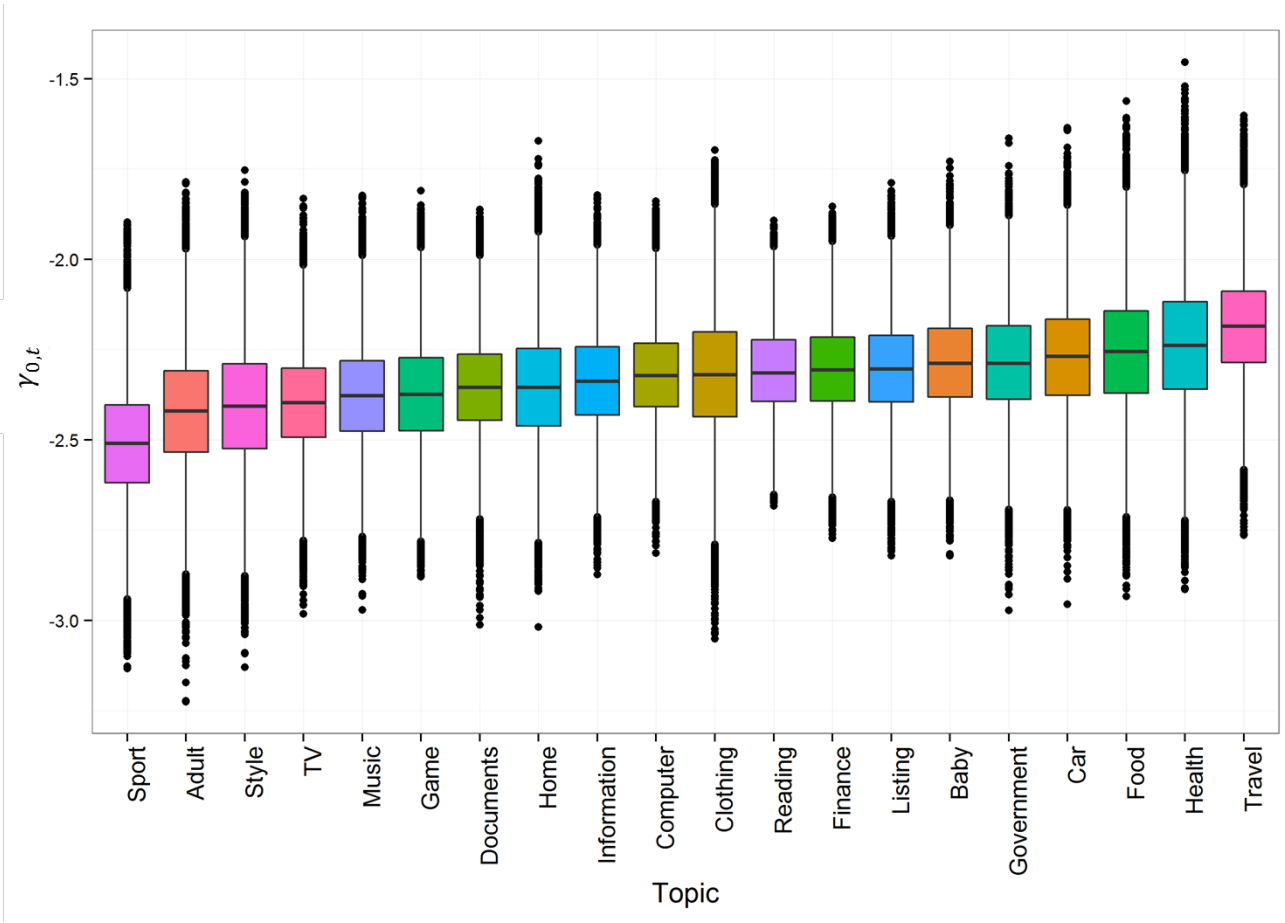

Figure A.2: Boxplots of Topic Specific Intercepts ($\gamma_{0,t}$)

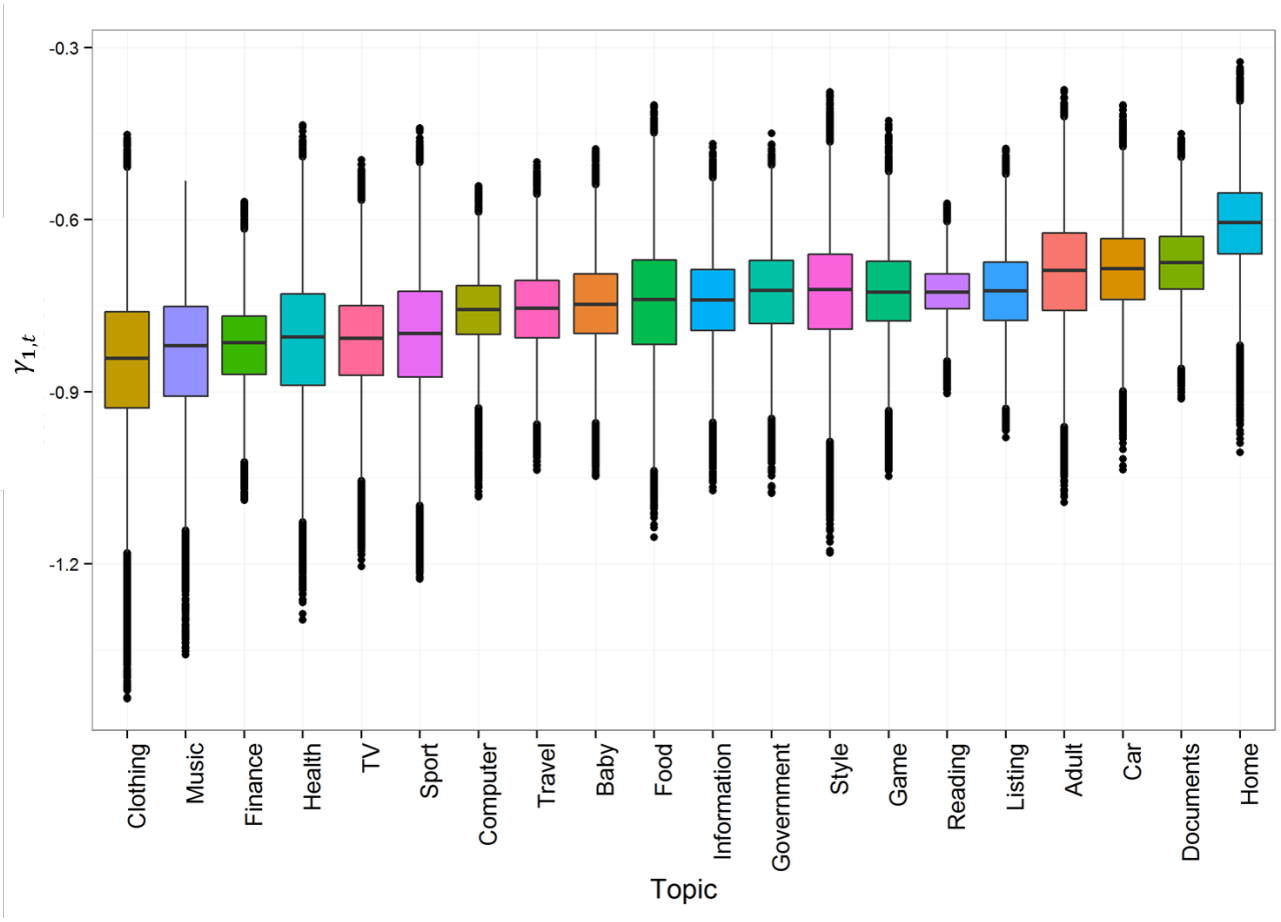

Figure A.3: Boxplots of Topic Specific Effects of Position ($\gamma_{1,t}$)

ONLINE APPENDIX J

NUMBER OF ORGANIC RESULTS FOR CORPUS CONSTRUCTION

We have compared the topic entropy values and empirical estimation results based on different numbers of Google results (i.e., top-50, top-60, top-80, and top-100), and present the comparisons below.

Comparing topic entropy. In the table below, we present the summary statistics for the computed topic entropy of the full data set (12,790 keywords) based on different numbers of organic search results. The high correlations among entropy values derived based on different numbers of organic search results suggest that entropy values seem to be fairly robust to the number of organic search results used to construct the corpus for topic modeling.

Table A.4: Entropy Values based on Different Number of Google Organic Search Results

	Mean	SD	Min	Max	Correlation			
					Top 50	Top 60	Top 80	Top 100
Top 50	1.60	0.45	0.34	2.99	1	0.87	0.86	0.85
Top 60	1.97	0.41	0.44	3.00	0.87	1	0.97	0.96
Top 80	1.99	0.40	0.44	3.00	0.86	0.97	1	0.97
Top 100	2.00	0.40	0.44	3.00	0.85	0.96	0.97	1

Comparing empirical results. We have further re-estimated the hierarchical Bayesian model using the entropy values and topic probabilities we now obtained based on different numbers of organic search results. We present the main results for CTR in the table below. As can be seen, the estimation results are fairly consistent across different columns, suggesting that our main results are robust to the number of organic search results used to cover the topics related to each keyword.

Table A.5: Estimation Results for CTR based on Different Number of Google Organic Search Results

	Variable	Top 50		Top 60		Top 80		Top 100	
Baseline (β_{0kt})	Intercept	-2.308***	(0.122)	-2.336***	(0.126)	-2.316***	(0.108)	-2.318***	(0.115)
	$TOPIC\_ENTROPY$	0.192***	(0.069)	0.223***	(0.074)	0.184**	(0.076)	0.165**	(0.072)
	NUM_WORDS	0.040	(0.045)	0.049	(0.045)	0.038	(0.046)	0.033	(0.045)
	$BRAND$	0.033	(0.064)	0.042	(0.064)	0.040	(0.064)	0.040	(0.064)
	$LOCATION$	-0.208**	(0.105)	-0.217**	(0.100)	-0.212**	(0.098)	-0.207**	(0.098)
	LOG_TRANS	0.147***	(0.029)	0.150***	(0.030)	0.152***	(0.029)	0.153***	(0.029)
	LOG_IMP	-0.010	(0.019)	-0.004	(0.020)	-0.004	(0.019)	-0.005	(0.019)
POS (β_{1kt})	Intercept	-0.726***	(0.045)	-0.749***	(0.066)	-0.727***	(0.059)	-0.724***	(0.055)
	$TOPIC\_ENTROPY$	-0.134***	(0.049)	-0.115**	(0.048)	-0.114**	(0.048)	-0.113**	(0.049)
	NUM_WORDS	-0.035	(0.032)	-0.035	(0.032)	-0.034	(0.030)	-0.034	(0.031)
	$BRAND$	-0.050	(0.045)	-0.056	(0.047)	-0.051	(0.044)	-0.047	(0.045)
	$LOCATION$	0.070	(0.064)	0.075	(0.065)	0.072	(0.063)	0.075	(0.063)
	LOG_TRANS	-0.014	(0.019)	-0.013	(0.019)	-0.015	(0.019)	-0.014	(0.019)
	LOG_IMP	-0.018	(0.013)	-0.019	(0.013)	-0.021*	(0.013)	-0.021	(0.013)
NUM_ADS (β_2)	NUM_ADS	0.166***	(0.017)	0.168***	(0.018)	0.167***	(0.015)	0.165***	(0.016)

***, **, and * indicate a 99%, 95%, and 90% significance level.