http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 37, No. 6, November–December 2018, pp. 930–952
ISSN 0732-2399 (print), ISSN 1526-548X (online)

A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries

Jia Liu,[a] **Olivier Toubia**[b]

[a] Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; [b] Columbia Business School, Columbia University, New York, New York 10025

Contact: jialiu@ust.hk, http://orcid.org/0000-0002-0279-724X (JL); ot2107@columbia.edu, http://orcid.org/0000-0001-7493-9641 (OT)

Received: December 15, 2015
Revised: February 28, 2017; February 6, 2018; April 23, 2018
Accepted: May 1, 2018
Published Online in Articles in Advance: October 16, 2018

https://doi.org/10.1287/mksc.2018.1112

Abstract. We extend latent Dirichlet allocation by introducing a topic model, hierarchically dual latent Dirichlet allocation (HDLDA), for contexts in which one type of document (e.g., search queries) are semantically related to another type of document (e.g., search results). In the context of online search engines, HDLDA identifies not only topics in short search queries and web pages, but also how the topics in search queries relate to the topics in the corresponding top search results. The output of HDLDA provides a basis for estimating consumers' content preferences on the fly from their search queries given a set of assumptions on how consumers translate their content preferences into search queries. We apply HDLDA and explore its use in the estimation of content preferences in two studies. The first is a lab experiment in which we manipulate participants' content preferences and observe the queries they formulate and their browsing behavior across different product categories. The second is a field study, which allows us to explore whether the content preferences estimated based on HDLDA may be used to explain and predict click-through rates in online search advertising.

History: K. Sudhir served as the editor-in-chief and Michel Wedel served as associate editor for this article.

Keywords: search engine optimization • search engine marketing • search queries • content preferences • semantic relationships • topic modeling

1. Introduction

Over the last decade, search engines, such as Google, have become one of the primary tools consumers use when searching for products, services, or information. This trend has given rise to two major industries, search engine optimization (SEO), whose related spending is expected to reach $80 billion by 2020 in the United States alone (Borrell 2016), and search engine marketing (SEM), whose related spending was estimated at $92 billion for 2017 (Statista 2017). SEO refers to the process of tailoring a website's content to optimize its organic ranking for a given set of keywords or queries to improve traffic and lead generation (Amerland 2013). SEM usually refers to paid advertising on search engines. Success in both of these industries hinges on firms' ability to infer the content preferences underlying consumers' search queries. For example, a firm engaging in SEM should bid more on keywords/queries that reflect preferences that are better aligned with its content. In addition, it should be able to identify search ad copies that optimally promote this content. Similarly, a firm engaging in SEO should promote its content, that is, attempt to have its content appear as a top organic search result to consumers for whom this content is more relevant. Hence, firms engaging in SEO should be able to assess which queries reflect content preferences that are best aligned with their content.

Despite the importance of being able to infer consumers' content preferences from their queries, very little research has been done in this area. Some research (which is reviewed in Section 2.2) has developed taxonomies of search queries and search intent. However, that research does not enable firms to infer content preferences in a quantified, nuanced, and detailed manner. Another stream of research in marketing (which is reviewed in Section 2.3) has quantified consumer preferences from their search behavior. However, that research has primarily focused on consumer search behavior that manifests itself via discrete choices (e.g., purchasing, clicking). Text-based search behavior (e.g., entering a search query), despite being a major way in which consumers search today, has not received as much attention in that literature.

Because of the nature of textual data, inferring content preferences from search queries presents several challenges. A first challenge is that search terms tend to be ambiguous; that is, consumers might use the same term in different ways. This implies that content preferences should be estimated taking into account the entire content of search queries. A second challenge is

the curse of dimensionality: the number of possible keywords or queries available to consumers is very large. A third challenge is the sparsity of search query: most search queries contain only up to five words (Wang et al. 2003, Kamvar and Baluja 2006, Jansen et al. 2009). Fourth, there exist some potentially complex semantic relationships between the content in a search query and the content in the corresponding search results. Previous research (also reviewed in Section 2.3) has suggested that consumers have the ability to leverage these semantic relationships when formulating queries. That is, consumers may not necessarily formulate queries that exactly and directly reflect their content preferences but rather formulate queries that are more likely to retrieve the type of content they are looking for.

The first two challenges may be addressed by simply describing content as a set of topics rather than individual words, following the literature in information retrieval (Manning et al. 2008). That is, content in queries and web pages may be described using a small number of topics, defined as probabilistic combinations of words. In this paper, we define a consumer's preferences as an ideal distribution across these topics, which reflects the content that the consumer wants to consume online. Such definition is analogous to the ideal-point model of preferences in which a product is preferred if it is closer to the consumer's ideal product profile (Green and Srinivasan 1978).[1]

However, addressing the third and fourth challenges calls for a different type of topic model that (1) is able to combine information from multiple sparse search queries and their associated search results and (2) explicitly quantifies the mapping between queries and results. We develop such a probabilistic topic model in this paper: hierarchically dual latent Dirichlet allocation (HDLDA). HDLDA is built upon latent Dirichlet allocation (LDA) (Blei et al. 2003), an unsupervised Bayesian learning algorithm that extracts "topics" from text based on occurrence. HDLDA is specifically designed for contexts in which one type of document (in our context, search queries) is semantically related to another type of document (in our context, web pages). The model is dual because the two types of document (search queries and web pages) share the same topic-word distributions. The model is hierarchical because the topic intensities of a web page are modeled as a function of the topic intensities of the search query(ies) that retrieve this page. Such structure alleviates the sparsity of search queries by allowing the topic intensities in a search query to be influenced by the information contained in the web pages that it retrieves as well as the other search queries that retrieve the same pages. Such structure also explicitly quantifies the mapping from search queries to search results. HDLDA can be estimated on any primary or secondary data set that contains the text of a set of queries and their results on a search engine.

HDLDA provides a basis for estimating consumers' content preferences from their queries. HDLDA models the topics in the web pages retrieved by a search engine in response to a search query; the model itself is agnostic as to how *consumers* translate their content preferences into search queries. Therefore, the exact manner in which content preferences are estimated based on HDLDA depends on the assumption the analyst is willing to make on how consumers translate their content preferences into search queries. If consumers are assumed to be strategic and formulate queries that will reach an ideal topic distribution among the results, their preferences may be estimated as the expected topic intensities of the results given their search queries. On the other hand, if consumers are assumed to be naive and formulate queries that directly express their content preferences, then their preferences may be estimated as the topic intensities of their search queries. In both cases, estimation may be made on the fly, making it useful for firms interested in customizing content (e.g., display or search advertising) based on a consumer's query.

We apply HDLDA and explore its use in the estimation of content preferences in two studies. We start by running a lab experiment that allows us to exogenously manipulate consumers' preferred content, which we can compare with estimates of content preferences. In particular, we provided participants with search tasks in various categories (e.g., finding a ski resort with specific features to recommend to someone) and asked them to perform a series of searches to find a suitable URL for each task description. To track user behavior on the search engine, we built our own search engine, Hoogle, which technically serves as a filter between Google and the user. Specifically, Hoogle runs all queries for all users through the Google application program interface (API), showing only the organic Google search results with no user history being captured (unlike with regular Google searches). In practice, marketers/advertisers may use a search engine's API to collect data on queries that are relevant for their search marketing strategies and use these data as input to HDLDA without running any experiment and without using a customized search tool, such as Hoogle. This is the approach we adopt in our second study. We illustrate the practical relevance of our research using field data collected by a large online travel company that heavily advertises on Google. We show field evidence that HDLDA may be used to explain and predict consumer click-through rate in online search advertising based on the degree of alignment between the search ad copy shown on the search engine results page and the content preferences estimated using HDLDA.

Our research has both methodological and managerial contributions. Methodologically, HDLDA has a structure that is different from current extensions of LDA in the marketing and topic-modeling literature. Managerially, HDLDA can help firms to organize and understand the meaning of different search queries and web pages and to estimate consumers' content preferences based on their queries. In general, by empowering marketers/advertisers to infer consumers' content preferences from queries, our work can help them create more relevant content and promote that content more effectively. In the context of SEM, our research can inform advertisers' bidding strategies by determining how well their content matches different queries. Our research can also help advertisers identify more relevant ad copy for a given search query, especially when there is not enough data on the click-through rate of each potential ad copy for that query or when advertisers have to manage thousands of possible ad copies and/or target queries (as is the case of the company with which we collaborated on our field study). In the context of SEO, our work can help firms prioritize their efforts by determining the queries on which it is more essential to improve organic search rankings, that is, the queries that reflect content preferences best aligned with the content they are trying to promote.

The rest of the paper is organized as follows. In Section 2, we review the related literature. In Section 3, we introduce the topic model HDLDA. In Section 4, we describe our experimental design and data. We present the results from our lab study in Section 5 and our field application in Section 6. We conclude in Section 7.

2. Relevant Literature

2.1. Natural Language Processing

There has been a stream of recent research in marketing that applies natural language processing (NLP) to analyze online user-generated content (Archak et al. 2011, Lee and Bradlow 2011, Ghose et al. 2012, Netzer et al. 2012). Our research builds upon the literature on topic modeling within NLP or the so-called LDA (Blei et al. 2003). LDA is an unsupervised Bayesian learning algorithm that extracts "topics" from text based on occurrence. By examining a set of documents, LDA represents each topic by a probability distribution over words and each document by a probability distribution over topics (to which we refer as topic intensities). Applications of LDA in the marketing literature include Tirunillai and Tellis (2014), who apply LDA to identify dimensions of quality and valence expressed in online reviews; Abhishek et al. (2018), who use LDA to measure the contextual ambiguity of a search keyword; and Büschken and Allenby (2017), who propose an extension of LDA in which words within the same sentence of an online review are constrained to pertain to the same topic.

Our proposed topic model, HDLDA, has a structure that differs from other extensions of LDA. We highlight three extensions related to ours: the correlated topic model (Blei and Lafferty 2007), the hierarchical topic model (Blei et al. 2003), and the relational topic model (Chang and Blei 2009). The correlated topic model allows correlation between documents in the occurrence of topics. In contrast, HDLDA focuses on the correlation between different types of document, for example, web pages and queries. The hierarchical topic model aims to learn a hierarchy of topics, that is, which topics are more general versus specific. In contrast, HDLDA defines hierarchy over documents; for example, topics in web pages are related to the topics in queries. The relational topic model studies document networks (e.g., whether two research papers tend to be cited by the same authors), whereas HDLDA leverages the observed hierarchy between different types of documents to infer their topics and semantic relationships.

2.2. Online Search Queries

Our topic model, HDLDA, captures the mapping between search queries submitted by users and search results provided by a search engine. This topic model allows researchers and practitioners to specify assumptions on how users translate their content preferences into search queries and develop methods for inferring content preferences from search queries given these assumptions and the mapping from search queries to search results provided by HDLDA. Hence, our work is relevant to the literature on understanding users' intent behind their search queries from the computer science and information systems literature. This research has primarily focused on classifying consumers' search intent into some discrete categories (Broder 2002, Jansen et al. 2007, Sanasam et al. 2008, Shen et al. 2011). The first and most popular categorization was proposed by Broder (2002), who defined three very broad classes: informational, navigational, and transactional. Informational search involves looking for a specific fact or topic, navigational search seeks to locate a specific website, and transactional search usually involves looking for information related to a particular product or service. Jansen et al. (2008) showed that about 80% of queries are informational, about 10% are navigational, and less than 10% are transactional.

Such empirical study of search logs provides valuable insights into what people search for and how they search for content. However, these types of analysis do not quantify users' content preferences, which HDLDA enables. This is managerially important to help website owners or advertisers improve the fit between their content and consumers' preferences.

2.3. Search Models

Consumer search behavior is often modeled within a utility maximization framework. Applications of this framework to online search have focused on discrete

search behavior, such as which links or products consumers decide to view/click. This was done either using static models or dynamic models in which users search sequentially and stop searching when the marginal cost of search exceeds the marginal gains (Jeziorski and Segal 2010, Kim et al. 2010, Dzyabura 2013, Ghose et al. 2013, Shi and Trusov 2013, Yang et al. 2015).

However, text-based search behavior, such as entering a search query, has been largely ignored. Entering a query is a first-order user behavior on most search platforms, and queries contain valuable information about user preferences (Pirolli 2007). Yet the field is lacking tools to leverage query data and in particular to extend search models based on utility maximization to the context of online search queries. Such extension requires specifying assumptions on how consumers formulate search queries given their content preferences. One such assumption was formulated by Liu and Toubia (2018), who argue that a query is not a direct representation of users' content preferences, but rather a tool to retrieve content that matches their preferences. These authors give the example of a consumer entering the following query: "affordable sedan made in America." It is possible that the most important attributes for this consumer are in fact safety, comfort, and made in America and that affordability is of lesser importance. This consumer might have decided to use this query because the consumer believes that cars made in America are generally safe and comfortable but not necessarily affordable. In that case, the consumer anticipated finding relevant search results efficiently (i.e., with short queries) by only including "made in America" and "affordable" in the queries but not "safe" or "comfortable" although these are important attributes. In other words, the consumer may have strategically leveraged the semantic relationships between queries and results when formulating the query. Liu and Toubia (2018) illustrate using field data that consumers stand to benefit from being strategic in query formation, and they present the results of an incentive-aligned lab experiment that suggests consumers have at least some ability to be strategic in query formation. Assuming that consumers are strategic in query formation leads to one particular way in which content preferences may be estimated from search queries, using the output of HDLDA. In this paper, we are agnostic ex ante as to how consumers translate content preferences into search queries. We empirically compare content preferences estimated using a strategic assumption to preferences estimated using a naive assumption that consumers formulate search queries that directly reflect their content preferences.

3. The Model

In this section, we first describe our proposed topic model, HDLDA, followed by its inference algorithm. Then, we show how the output from HDLDA can be used to estimate consumer content preferences based on search queries. We also present some benchmark approaches, which we compare with HDLDA in our empirical studies.

3.1. HDLDA

HDLDA is a model for bag-of-word data with which one type of document—in our case, search queries—is semantically related to a different type of document—in our case, web pages. We assume that there is one LDA process for each type of document. The two processes share the same topic-word distributions, and they are hierarchical in the sense that the topic intensities in each web page are related to the topic intensities in the query(ies) that retrieve the page. HDLDA can be applied to any corpus that has such hierarchically dual structure. We focus here on an application to search engines and set the notations within this context.

Suppose there is a collection of Q different queries for a particular search domain, and these queries retrieve a collection of P different web pages on a search engine. Let $l_{pq} \in \{0,1\}$ indicate whether web page p is retrieved by query q; that is, it appears in the top search results for query q. Let J denote the total number of different words in the vocabulary, and words are indexed by $j \in \{1,2,\ldots,J\}$. Let w_{qj} denote the jth word in query q and w_q denote the vector of J_q words associated with that query, where J_q is the number of words in the query. Similarly, let w_{pi} denote the ith word in web page p and w_p denote the vector of J_p words associated with that web page, where J_p is the number of words in the page. The set of relationships between the data and the model parameters is described by the graphical model in Figure 1. Note that we treat the labels $\{l_{pq}\}$ as exogenously given by the search engine; that is, we do not model their generating process.

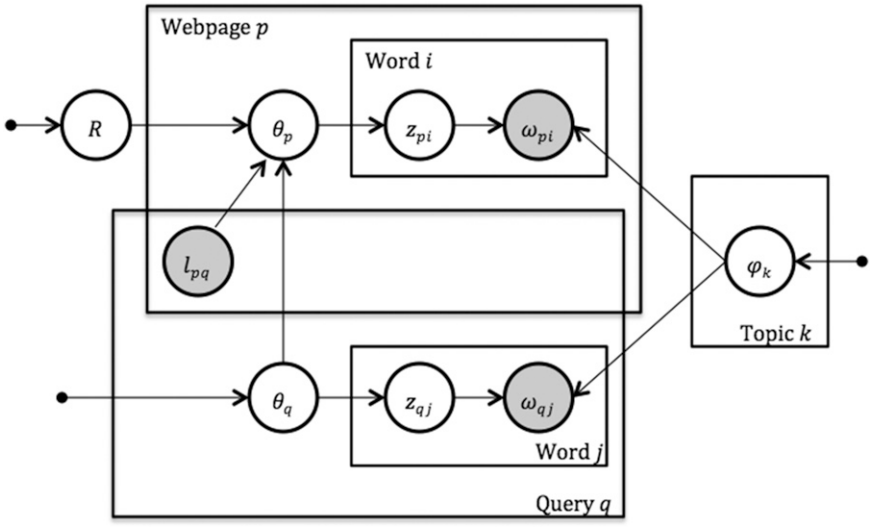
Topics. We let K denote the number of different topics in the domain. Search queries and web pages in the collection are assumed to share the same set of topics, but each document exhibits these topics with different intensities. These topic intensities are reflected by the words present in the documents. Similarly to an LDA, we model each topic $k \in \{1,2,\ldots,K\}$ as a vector φ_k, which follows a Dirichlet distribution over the J words in the vocabulary:

$$\varphi_k \sim Dirichlet_J(\eta). \tag{1}$$

The hyper-parameter η is a scalar that we estimate, which controls the sparsity of the word distribution.

Queries. To model the observed jth word w_{qj} in each query q, we need to model the query's topic intensities,

Figure 1. The Graphical Model of HDLDA



captured by the vector θ_q and the latent topic assignment $z_{qj} \in \{1, 2, \ldots, K\}$ for that word. Following LDA, we assume for $q = 1, 2, \ldots, Q$ and $j = 1, 2, .., J_q$

$$\theta_q \sim Dirichlet_K(\alpha),\ z_{qj} \sim Category(\theta_q),\ w_{qj} \sim Category(\varphi_{z_{qj}}). \tag{2}$$

The hyper-parameter α is a scalar that controls the prior on the topic intensities in the queries, which we set to a fixed value in this paper.

Web Pages. Web pages are semantically related with the set of queries that can retrieve them. HDLDA captures such a relationship by incorporating a hierarchical structure. Specifically, we model the prior on the topic intensities for web page p, θ_p, as a function of the topic intensities of the queries that retrieve this web page. The mapping between queries and results is specified at the topic level, using a $K \times K$ matrix R. In this matrix, each element $r_{kk'}$ indicates the effect of topic k in the retrieving queries on topic k' in the corresponding search results. As multiple queries may retrieve the same web page, we use the average topic intensities across these queries, which is denoted as $\overline{\theta}_q(p) = \sum_q \theta_q l_{pq} / \sum_q l_{pq}$ in the following equation.[2] Following the Dirichlet-multinomial regression topic model (Mimno and McCallum 2008), we assume that

$$\theta_p \sim Dirichlet_K(\exp(R_1^T \overline{\theta}_q(p)), \ldots, \exp(R_K^T \overline{\theta}_q(p))). \tag{3}$$

The exponential of the product between the kth column of R and $\overline{\theta}_q(p)$ is proportional to the expected intensity of topic k in the search results. That is, the intensity of each topic in each document is related to the intensities of all topics in the labeling query(ies). Given θ_p, the observed jth word in web page p is then modeled in a standard manner

$$z_{pj} \sim Category(\theta_p),\ w_{pj} \sim Category(\varphi_{z_{pj}}) \tag{4}$$

for $p = 1, 2, \ldots, P$ and $j = 1, 2, .., J_p$.

3.2. Inference Algorithm

Given the content of all queries and web pages and the labeling of web pages by queries, our goal is to estimate $\Theta = \{\{\varphi_k\}, \{z_p\}, \{z_q\}, \{\theta_p\}, \{\theta_q\}, R\}$. We use a combination of Markov chain Monte Carlo (MCMC) and optimization, that is, a stochastic expectation maximization (EM) sampling scheme (Diebolt and Ip 1995, Nielsen 2000, Mimno and McCallum 2008). Specifically, we apply a Gibbs sampler to draw $\{\varphi_k\}, \{z_p\}, \{z_q\}, \{\theta_p\}$ from their posterior distributions, which are all conjugate; we use the Metropolis–Hastings algorithm to sample $\{\theta_q\}$, which are not conjugate; and we estimate R by maximizing its likelihood function given $\{\theta_q\}$ and $\{\theta_p\}$. Therefore, over the MCMC iterations, we alternate between sampling $\{\{\varphi_k\}, \{z_p\}, \{z_q\}, \{\theta_p\}, \{\theta_q\}\}$ and numerically optimizing R given the other parameters.[3] The details of our inference algorithm are presented in Appendix A. In Appendix B, we report a simulation study that explores the performance of this algorithm.

Hyper-Parameters. The extant literature suggests that optimizing the hyper-parameters may improve the performance of a topic model (Wallach et al. 2009a, b). We tried to estimate both α and η using the previously stated algorithm by optimizing their likelihood functions, respectively. However, we found consistently across multiple corpora that these two hyper-parameters cannot be jointly estimated in this application, which we find is due to the sparsity of search queries. Therefore, we

set $\alpha = 0.1$, and we estimate η by maximizing its likelihood function given $\{\varphi\}$. We treat η as a scalar (giving rise to a symmetric prior) as Wallach et al. (2009a, b) find that an asymmetric prior over the topic-word distributions provides no real benefit.[4]

3.3. Estimating Content Preferences Based on Queries

HDLDA is a topic model that relates the content in search queries to the content in the web pages retrieved by a search engine in response to these queries. This model in itself is agnostic as to how consumers translate their content preferences into search queries. Nevertheless, HDLDA allows practitioners and researchers to specify an assumption on how users translate their content preferences into search queries and then use the model to infer or reverse engineer content preferences from search queries. In this paper, we consider two alternative assumptions on how users translate their content preferences into search queries, which give rise to two alternative estimation approaches.

The first assumption (consistent with Liu and Toubia 2018) is that consumers anticipate the types of results that will be retrieved by their query and that they formulate queries that will retrieve results that match their preferred content in expectation. We label this assumption "strategic" because it assumes that users strategically leverage semantic relationships in query formation. Under this assumption, a consumer's preferences may be estimated as the expected topic intensities in the search results given their query.

The alternative assumption we consider is that users do not leverage the semantic relationship between queries and results when formulating their queries. That is, users formulate queries that directly reflect their content preferences rather than formulating queries that will retrieve results that reflect these preferences. We label this assumption "naive" because it assumes consumers ignore the mapping between queries and results. Under this assumption, a consumer's preferences may be estimated directly as the topic intensities in the search queries. We note that these two assumptions constitute the two ends of a continuum of possible assumptions that would allow users to have only approximate beliefs on the relevant semantic relationships and/or an imperfect ability to leverage these relationships. We leave the testing of such assumptions to future research.

We define consumer i's content preferences as an ideal distribution over topics on web pages. This distribution is captured by a vector of weights across K topics, denoted as β_i. Suppose we observe query q from consumer i. Given the topics $\{\varphi\}$ already estimated from HDLDA, we run an LDA to obtain an estimate of the topic intensities of query q, denoted as $\widehat{\theta}_q$. According to HDLDA, the search engine will retrieve web pages whose topic intensities are drawn from the following distribution: $\theta_p \sim Dirichlet_K(\exp(R_1^T\widehat{\theta}_q), \ldots, \exp(R_K^T\widehat{\theta}_q))$. Accordingly, under the strategic assumption, β_i may be estimated as the mean of the expected topic intensities in search results:[5]

$$\begin{aligned}\widehat{\beta}_i^{HDLDA_{strategic}} &= E(\theta_p \mid \widehat{\theta}_q) \\ &\triangleq \left(\frac{\exp(R_1^T\widehat{\theta}_q)}{\sum_k \exp(R_k^T\widehat{\theta}_q)}, \frac{\exp(R_2^T\widehat{\theta}_q)}{\sum_k \exp(R_k^T\widehat{\theta}_q)}, \ldots, \frac{\exp(R_K^T\widehat{\theta}_q)}{\sum_k \exp(R_k^T\widehat{\theta}_q)}\right). \quad (5)\end{aligned}$$

In contrast, under the naive assumption, β_i may be estimated as the topic intensities of the search query itself:

$$\widehat{\beta}_i^{HDLDA_{naive}} = \widehat{\theta}_q. \quad (6)$$

Benchmark. We compare the estimation of content preferences using the output of HDLDA to a benchmark in which content preferences are estimated based on LDA, which treats queries and web pages as independent documents. We ensure that the comparisons of LDA to HDLDA not be driven by HDLDA having a flexible prior distribution on θ_p. Specifically, we allow LDA to also have a flexible prior, $\theta_p \sim Dirichlet_K(\alpha_{page})$, and we estimate the $1 \times K$ vector of asymmetric hyper-parameters α_{page}. Similarly to HDLDA, we set $\theta_q \sim Dirichlet_K(\alpha)$ as the prior on the topic intensities in queries, where $\alpha = 0.1$, and $\varphi \sim Dirichlet_J(\eta)$ as the prior on the topic distributions, where η is a scalar that we estimate. In this case, as semantic relationships are not captured, content preferences are estimated as the estimated topic intensities of the query:

$$\widehat{\beta}_i^{LDA} = \widehat{\theta}_q. \quad (7)$$

This benchmark is nested within HDLDA in which we assume all the topics in a query have the same effect on each topic in the results; that is, $R = (r_1\mathbb{1}_K, r_2\mathbb{1}_K, \ldots, r_K\mathbb{1}_K)$, where $\mathbb{1}_K$ denotes a K-dimensional vector of ones and $\{r_k\}_{k=1,\ldots,K}$ are all scalars. This reduces the mean of the Dirichlet distribution in Equation (3) to $\alpha_{page} = (\exp(r_1), \exp(r_2), \ldots, \exp(r_K))$. Similarly to HDLDA, we estimate this benchmark with a stochastic EM algorithm in which we alternate between sampling $\{\{\varphi_k\}, \{z_p\}, \{z_q\}, \{\theta_p\}, \{\theta_q\}\}$ from the Gibbs sampler and numerically optimizing the prior parameters α_{page} and η given the other parameters.

4. Lab Experiment

Researchers and practitioners may run HDLDA on any primary or secondary data set that contains the text of a set of queries and their corresponding search results

from a search engine. However, in this paper, our objective is not only to show how HDLDA may be applied, but also to test the model's usefulness as a basis for inferring consumers' content preferences from their search queries. Accordingly, our first data set was collected experimentally, which enabled us to manipulate the content preferences underlying users' search behavior and measure (albeit imperfectly) the users' "true" content preferences, which we then compared with various estimates. To track user behavior on the search engine, we built our own search engine, Hoogle, which technically serves as a filter between Google and the user. Hoogle has the additional benefit of removing the influence of advertising and customization on user search behavior, therefore providing clean comparisons between benchmarks. In this section, we describe this data set in detail.

4.1. Experimental Design

We conducted a lab experiment in which $N = 197$ participants performed a series of online search tasks using our custom-built search engine Hoogle, which we introduce in Section 4.2. The participants' objective was to make purchase recommendations. Our search tasks were designed based on five product categories about which consumers commonly acquire information on the internet before purchase: ski resorts, printers, cars, laptops, and cameras.[6] We manipulated content preferences exogenously by giving participants specific search tasks, that is, descriptions of what they should search for. Each participant was asked to submit one URL of their chosen web page that they believed best matched the given task description. To ensure that participants' preferences were aligned with the task descriptions and their corresponding chosen web pages, our study was incentive-aligned. Participants were told that all submitted links would be evaluated by the researchers based on relevance and usefulness. In addition to a $7 participation fee, we gave a $100 cash bonus to the participant whose chosen web page best matched with the corresponding task description. Participants were informed of this incentive before the experiment, and we notified the winner within two weeks of the experiment.

We introduced some heterogeneity in content preferences by designing two task descriptions reflecting different preferences within each category as displayed in Table 1. For example, task 1 asked participants to search for a family-friendly ski resort, and task 2 asked them to search for an exclusive ski resort.[7] We used a between-subjects design in which each participant was randomly assigned to one version of the two tasks in each category. The order of the categories was randomized for each participant. We find participants spent, on average, 20 minutes to finish all the tasks, which suggests that our incentives worked well.

4.2. Data Collection

As mentioned previously, in addition to applying HDLDA in various domains, our objective with the present experiment is also to explore its use for inferring a user's preferences based on their search queries. To track user behavior on the search engine while ensuring that our comparison of various benchmarks not be influenced by unobserved factors, such as the user's browsing history or the customization of content by the search engine, we built our own search engine called Hoogle and used it to collect search queries from consumers. Hoogle retrieves all the organic search results for each search query with no user history being captured, using the Google customer engine API. That is, for any search query, Hoogle retrieves a similar set of search results as Google with the only differences that search results are not personalized based on past search history and there is no sponsored search result. A screenshot from the Hoogle interface is presented in Figure 2. Each result page shows 10 links with their titles and snippets. The font, color, and size are the same as Google.

The search logs from Hoogle include the following information for each participant and for each task: the query(ies) submitted by the participant, the search results seen by the participant on each page, and the links clicked by the participants. Immediately after we finished collecting data in the lab experiment, we also used Python scripts to automatically download all the content of the web pages in the participants' search results (i.e., the actual content of the web pages corresponding to all the links in the result pages viewed by participants).

We note again that Hoogle is not necessary to run HDLDA. In practice, most firms have a set of keywords/queries that they think are most relevant and valuable for their SEO/SEM strategies. For example, in our field application, the set of consumer queries was collected based on a subset of the keywords on which the firm frequently advertises. We also note that Hoogle is based on the Google API; that is, the organic results associated with each search query come directly and only from Google even though the set of queries comes from the interaction of consumers with Hoogle.

4.3. Descriptive Statistics

Table 2 reports descriptive statistics on the search data collected from this lab experiment for each category, including the number of unique queries, the number of different words among all the queries, the variation across users' queries, users' query usage, the number of words per query, and the average proportion of the words in a query that come from the task description. First of all, there exists some heterogeneity across users' queries. Such variation is measured by the edit distance (Jurafsky and Martin 2000), which is a way to quantify

Table 1. Search Tasks

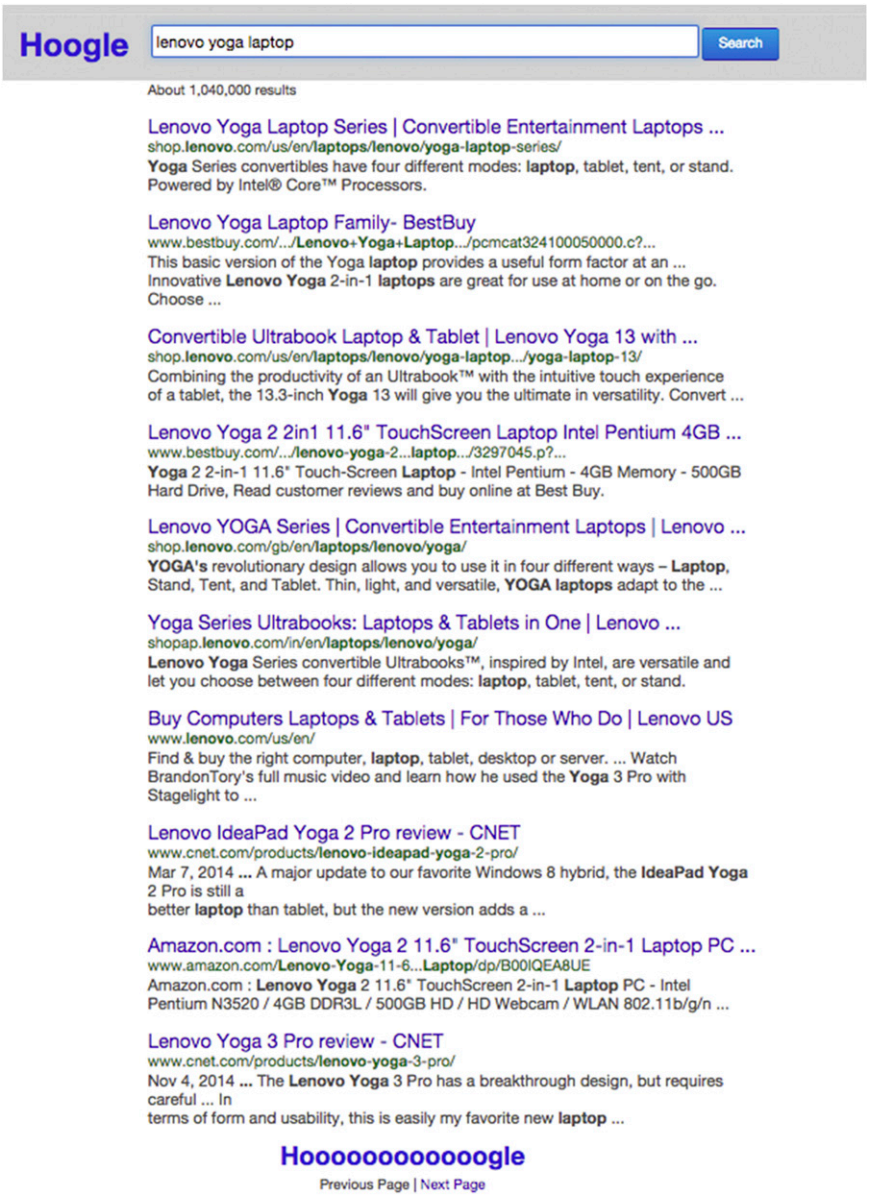
Task number	Task description
Ski	
1	A family is planing to take a vacation at a ski resort in Vermont. They are looking for a small resort that is suitable for family and children. The resort should have plenty of trails for beginner skiers and also a ski school for kids. There should be lesson package deals including all-day lift tickets, ski rental, lessons, and so on. At the minimum, the resort should have an on-site ski rental shop and offer some kind of discounts.
2	David, a banker, is planning to take a vacation at a ski resort in Vermont. He is looking for an exclusive resort that could offer a variety of terrains for intermediate and advanced skiers. Specifically, the mountain should be large, and the slopes should be somewhat difficult. In addition, the resort should offer other activities, such as snow tubing, and amenities, such as a lounge and spa treatments.
Printer	
3	Jessica, a college student, wants to purchase a budget printer for school work. The printer should be able to print, copy, and scan. Double-sided printing would also be attractive to her. Color printing is not required as she will mostly print black and white. In addition, the printer should print fast with low noise and the running cost should be low.
4	A family wants to purchase a small printer designed for home users who want lab-quality photos. They want to be able to connect the printer to wi-fi and smart phones, and it should also be able to print photos without the topic of a computer. As the printer will be used very frequently, the family is willing to pay slightly more for a printer that is cheaper to run in the long term.
Car	
5	A family wants to buy a new car that could provide more generous space for seating and cargo than their old compact sedan. The family's budget is $25,000. They want the new car to be safe, reliable, economic, and fuel efficient. It should have a four-cylinder engine and a high EPA mileage. The car should also handle snow and ice well.
6	Catherine wants to buy a small car to save money on gas, insurance, and maintenance. She also wants to be able to park more easily in a big city. Her price range is between $10,000 and $14,000. She wants the car to be attractive, stylish, fun, and practical. Despite its small size, the car should still be safe and should offer a comfortable ride.
Laptop	
7	Mike, a college student, wants to buy a new laptop. In addition to school work, the laptop should provide good performance for playing games. It should have at least an Intel Core i5 CPU, 8 GB of RAM, a good graphics card, and a larger screen. Mike prefers windows and Linus systems because of their flexibility and wide options for programs. Mike's price range is between $800 and $1,000.
8	Mike, a consultant, wants to buy a laptop for work and traveling. Mike's budget is $600. He will mostly use the laptop for internet, Word, PowerPoint, and email. He needs a laptop with enough speed, good display, very long battery life, small size, and light weight. Also the laptop should be durable enough to handle pressure or dropping that may often happen during traveling.
Camera	
9	A couple wants to buy a camera for their nine-year-old son. The camera should be simple to use and easy to carry anywhere. The camera should have a large viewing screen or touchscreen. Its picture quality should be very good. More importantly, the camera should be sturdy and be able to withstand falls. And it should also be waterproof, so that it can be used underwater. The couple prefers a camera in the price range of $100–$200.
10	Kevin, a beginner photographer, wants to buy his first digital SLR camera. Kevin is looking for a model in the midprice range or alternatively a package kit that includes the body, lenses, and tripod. The camera should be able to shoot both jpeg and raw files. And it should also come with a wi-fi adapter, which makes it easier to quickly share images through a laptop or phone.

how dissimilar two strings are to one another.[8] We compute the edit distance between all possible pairs of queries from users in the same task and report the sample average and standard deviation. In addition, users' query usage varies across categories, but on average, they tend to form few queries in a session and use few words in a query, which is consistent with the existing empirical studies on search queries (Jansen et al. 2000, Kamvar and Baluja 2006, Wu et al. 2014). Finally, the average proportion of words in users' search queries that are from the task description ranges from 0.40 to 0.75. Hence, users form queries that combine words in the task description with other words.

We now turn to descriptive statistics related to position effects. Previous research has shown that position could affect behavioral outcomes, such as consumer click-through, conversion rate, and sales (Kihlstrom and Riordan 1984, Hotchkiss et al. 2005, Varian 2007). In our data, we find that the majority of users limit their browsing to the links on the first page of results (i.e., the top 10 organic links retrieved by Google). Therefore, we treat every page of results as starting from the first position. We plot the click-through rate (CTR) against the top 10 positions in Figure 3. The CTR at each position is calculated as the percentage of clicks at this position across all the clicks from the 10 positions, and hence, the CTR sums to one across positions. Consistent with previous research, we see a quasi-exponential decrease in CTR as a function of position (Narayanan and Kalyanam 2015, Abhishek et al. 2018).

Finally, we study whether there exists some agreement on the best web page among users assigned to the

Figure 2. (Color online) The Interface of Hoogle



same task. For each task and for each URL that was chosen at least once, we define the agreement level as the proportion of users who picked that URL as their chosen link for that task. Overall, the distribution of the agreement level is long-tailed; that is, the majority of users tend to choose different URLs. We report in Figure 4 the agreement levels for the top five most chosen links for each task. We see that a few links show some level of consensus among users, and there exists heterogeneity across tasks.

4.4. Data Preprocessing

Given that most advertisers or firms do business in certain domains and design web page content and keyword lists within that, we run HDLDA separately for each product category. Accordingly, we combine all the queries and web pages from the two search tasks in the same category as one corpus. We preprocess the text in each corpus based on standard practice in text mining. We remove any delimiting character, including hyphens; we eliminate punctuation, non-English characters, and a standard list of English stop words; no stemming is performed. We form the vocabulary for each corpus using the standard term frequency-inverse document frequency metric (Jurafsky and Martin 2009).[9]

The descriptive statistics of the resulting corpora are summarized in Table 3. The first row is the number of words that are selected as the vocabulary for each corpus. The number of words in each query or web page is calculated based on the selected vocabulary rather than the original content. Hence, one may notice that these numbers are smaller than those in Table 2.

Table 2. Descriptive Statistics of Users' Search Queries

	Task	Number of users	Number of unique queries	Number of unique words	Edit distance	Number of queries per user	Number of words per query	Overlap with description
Ski	1	99	173	111	31.41 (21.35)	2.13 (1.56)	5.36 (2.62)	0.76 (0.25)
	2	97	187	118	33.26 (26.52)	2.54 (1.94)	5.22 (3.14)	0.75 (0.28)
Printer	3	97	295	183	32.56 (14.73)	3.47 (2.93)	5.93 (3.14)	0.57 (0.32)
	4	97	204	162	30.82 (25.94)	2.51 (2.02)	4.67 (2.19)	0.53 (0.30)
Car	5	97	375	262	35.71 (20.61)	4.68 (2.93)	5.91 (3.31)	0.56 (0.36)
	6	99	368	244	24.53 (16.86)	2.51 (2.02)	4.15 (1.88)	0.42 (0.36)
Laptop	7	98	338	243	32.12 (15.36)	4.06 (4.17)	6.43 (3.29)	0.58 (0.36)
	8	95	292	250	30.54 (26.90)	3.67 (3.03)	4.34 (2.27)	0.43 (0.37)
Camera	9	98	243	214	31.13 (18.89)	2.95 (2.51)	4.71 (2.22)	0.44 (0.29)
	10	95	271	141	28.65 (16.61)	3.66 (2.78)	5.28 (2.81)	0.68 (0.34)

Notes. We report the sample average with the standard deviation in parentheses. The edit distance is the minimum number of operations required to transform one query into the other. Larger values indicate lower similarity. We compute the edit distance between all pairs of queries (queries are pooled together across users) for the same task. The last column reports the proportion of words in a user's query that appear in the task description.

On average, each query contains about four words, and each web page contains 250–600 words. The last part of the table concerns the observed labels between queries and web pages. On average, each query retrieves about 10 web pages.[10] Web pages can be retrieved by very different numbers of queries.

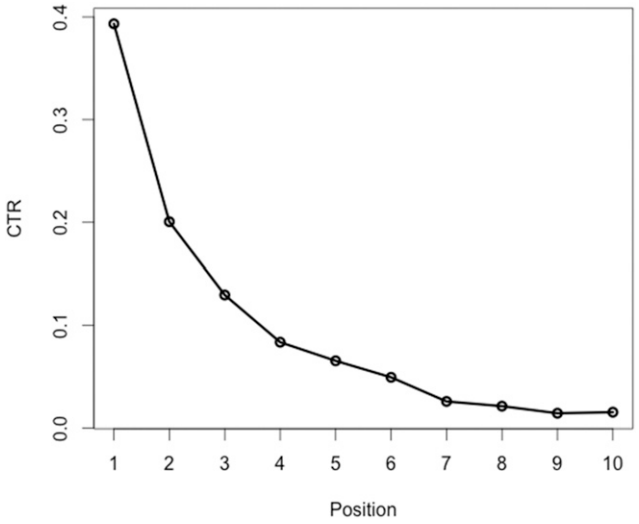
5. Lab Experiment Results

In this section, we first describe model estimation. Next, we present the results of the posterior estimates from HDLDA. Finally, we proceed to the individual-level estimation of content preferences as described in Section 3.3.

5.1. Model Estimation

One key decision in topic modeling is choosing the number of topics for a corpus if this parameter is not specified a priori (Blei and Lafferty 2009). Depending on the goals and available means, a researcher may apply a variety of performance metrics (Griffiths and Steyvers 2004; Chang and Blei 2009; Wallach et al. 2009a, b). In our case, we initially intended to use the number of topics determined by evaluation on hold-out documents for the benchmark model LDA described in Section 3.3 as the number of topics for HDLDA. However, we found that LDA prefers very large K.[11] This issue has been documented in other empirical applications in marketing (Trusov et al. 2016, Zhang et al. 2017). Moreover, Chang et al. (2009) found that topic models that perform better on held out likelihood (e.g., measured by perplexity) may infer less semantically meaningful topics. Therefore, we set $K \in \{2, 3, 4\}$ for each corpus based on interpretability. We also estimate all the benchmark models for $K \in \{2, 3, 4\}$ to evaluate the robustness of our results to different choices of K.[12] We hope that future research will propose more effective, objective, and systematic methods for determining the optimal number of topics in HDLDA and other topic models. We compare the model fit of HDLDA and LDA based on the deviance information criterion (DIC) (Spiegelhalter et al. 2002). The results are reported in Table 4. We find consistently, across all the categories and K, that HDLDA achieves a much lower DIC compared with LDA. This suggests that it is reasonable to explicitly model the mapping between search queries and search results.

Figure 3. CTR as a Function of Position



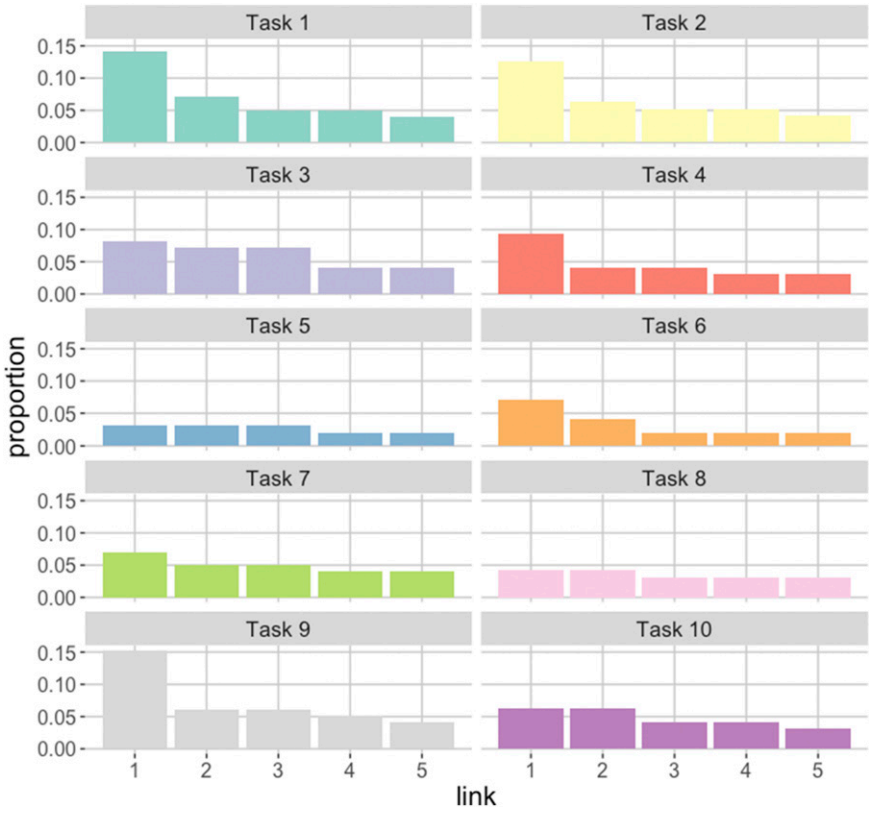
Note. CTR is normalized to sum to one across positions.

5.2. Posterior Estimates

We now interpret the topics generated by HDLDA in each corpus. To ease interpretation, we focus on the most relevant words in each topic. The relevance of word w to topic k is measured as follows (Bischof and Airoldi 2012, Sievert and Shirley 2014):

$$r(w, k \mid \lambda) = \lambda \log(\varphi_{kw}) + (1 - \lambda) \log\left(\frac{\varphi_{kw}}{p_w}\right), \quad (8)$$

Figure 4. (Color online) Agreement Level of the Top Five Most Chosen Links for Each Task



Note. For each URL, its agreement level is defined as the proportion of users who picked that URL as their chosen link for the same task.

where φ_{kw} is the posterior estimate of the probability of seeing word w given topic k, p_w is the empirical distribution of word w in the corpus, and λ determines the weight given to the probability of word w under topic k relative to its lift $\frac{\varphi_{kw}}{p_w}$, both measured on the log scale. Setting $\lambda = 1$ results in the familiar ranking of words in decreasing order of their topic-specific probabilities, and setting $\lambda = 0$ ranks words solely based on lift. We set $\lambda = 0.6$, following the empirical studies conducted by Sievert and Shirley (2014).

For ease of interpretation, we simulate the content of each topic using the exponential of relevance. That is, we generate sets of words for each topic, in which the probability of occurrence of each word is proportional to the exponential of its relevance. We use word clouds to visualize the simulated sets of words. As an example, in Figure 5, we report the word clouds for the four topics extracted from the laptop category when setting $K = 4$. Words with larger font size have higher relevance. Based on the word clouds in Figure 5, one may label topic 1 as "shopping for laptops," topic 2 as "Lenovo related," topic 3 as "performance," and topic 4 as "configuration."

After examining all the extracted topics, we set $K = 2$ for ski and camera, $K = 3$ for printer and car, and $K = 4$ for laptop. Table 5 displays some of the most relevant words for each topic in each category along with examples of queries and web pages with very high

Table 3. Descriptive Statistics of the Corpora

	Ski	Printer	Car	Laptop	Camera
Vocabulary size	1,709	3,258	3,128	3,321	3,309
Unique queries	351	495	749	631	515
Words per query	4.62 (2.31)	3.84 (2.01)	3.83 (2.14)	4.50 (2.87)	4.14 (2.03)
Unique web pages	1,167	1,984	4,253	3,042	2,238
Words per web page	258 (278)	366 (341)	559 (697)	736 (1862)	444 (525)
Query–page pairs	3,636	4,928	7,851	6,454	5,049
Web pages per query	10.36 (1.89)	9.96 (2.00)	10.48 (5.61)	10.23 (2.82)	9.80 (2.96)
Queries per web page	3.12 (6.51)	2.48 (4.41)	1.85 (2.72)	2.12 (4.22)	2.26 (3.84)

Note. We report the average across all participants with standard deviations in parentheses.

Table 4. DIC

Model	*K*	Ski	Printer	Car	Laptop	Camera
HDLDA	2	3,122,672	8,236,674	27,683,535	25,298,079	11,261,204
	3	2,973,582	7,893,707	26,529,830	24,384,996	10,855,606
	4	2,849,094	7,650,230	25,744,413	23,963,219	10,456,815
LDA	2	3,165,124	8,352,227	28,017,795	25,825,592	11,375,783
	3	3,031,919	8,081,503	27,298,308	25,080,194	11,084,656
	4	2,924,902	7,867,257	26,704,227	24,665,402	10,792,368

weights on each topic. Although we only present five sample words per topic for space reasons, as one may see in Figure 5, this is not enough to define or capture a topic completely. We observe that the recovered topic intensities of web pages in general seem to be consistent with their actual content. Table 6 reports the average of the posterior estimates of the topic intensities of all queries and web pages in each category. We see that the average θ_q may not necessarily be similar to the average θ_p.

5.3. Content Preference Estimation

Measures of "True" Content Preferences. One of the appeals of our experimental design is that we can manipulate content preferences exogenously by explicitly instructing participants to search for certain content. We do not claim to be able to measure with certainty how participants interpreted the task and, therefore, what their true underlying content preferences were during the experiment. Nevertheless, our experimental design provides us with some (imperfect) measures of participants' true content preferences, which we may then compare with the content preferences estimated from participants' queries based on various approaches. Our first measure of "truth" is the set of topic intensities of the actual web page chosen by the participant.[13] This measure has the benefit of reflecting the actual behavior of participants. However, one limitation of this measure is that a particular page is more likely to be chosen if it is one of the top search results, and HDLDA precisely models the expected distribution of topics across top search results given a query. That is, the chosen web page is not only influenced by content preferences (i.e., the demand side), but also by the options presented by the search engine (i.e., the supply side), and hence, this measure partly reflects how well the various models capture the supply side. So we complement this with a second measure of true preferences: the set of topic intensities of the task description given to the participant. This measure offers the benefit of being unaffected by the options presented to the participants by the search engine. However, one possible drawback of this measure is that there might be variations in how participants interpret the task description. Note that the data used to train HDLDA contains neither the text of the task descriptions nor the knowledge of which web page was selected by each participant. Therefore, both our truth measures may be viewed as external validations.

Before comparing performance across benchmarks, we provide some additional statistics on our truth measures. The estimated topic intensities of all the task descriptions are presented in Table 7.[14] We see that each task description may have large intensities on multiple topics. However, when comparing the intensities of the same topic across the two task descriptions in the same category, the topic with the relatively larger intensity is consistent with our expectation. For example, in the ski resort category, task 1 (respectively, task 2) was designed to correspond to a family-friendly resort (respectively, a luxury resort). Consistent with this, we find that task 1 has a larger intensity on topic 1 compared with task 2, and the opposite is true for topic 2. Note however that the intensities on topic 1 are larger overall compared with the intensities on topic 2, reflecting the fact that family-friendliness is a more common/popular theme in this category compared with luxury.

Finally, for each subject in each task, we compare the topic intensities of the web page chosen by the subject with the topic intensities of the links that the subject clicked on but did not choose and of all links displayed on the search engine result page for that subject. The similarity of two sets of topic intensities is measured using cosine similarity (i.e., inner product between two vectors),[15] which is commonly used in topic modeling to understand the similarity between documents. In our case, it ranges from zero, indicating complete orthogonality, to one, meaning perfect alignment. We report the results across all K and product categories separately for different topic models in Appendix C. We see that the similarity between the content that participants end up choosing and the content on which they tend to click is greater than the similarity between the content they end up choosing and the content on any search engine result page.

Performance Metric. As a performance metric, we compute the *perplexity* score of the true description of each participant's content preferences (i.e., task description or chosen web page) given the estimated content preferences. Perplexity is monotonically decreasing in the likelihood of the data and is equivalent

Figure 5. (Color online) Word Cloud of Four Topics in Laptop

Notes. Panels (a)–(d) are the simulated content for topics 1–4 in the laptop category. The size of each word is proportional to the exponential of its relevance.

to the inverse of the geometric mean of the per-word likelihood (Blei et al. 2003). A lower perplexity score indicates better model performance. Let L_i denote the content (i.e., set of words) of the true description of user i's preferences and $|L_i|$ denote its length. The perplexity score of L_i given a vector of estimated content preferences $\widehat{\beta}_i$ is calculated as

$$perplexity(L_i|\widehat{\beta}_i) = \exp\left(-\frac{\sum_{w\in L_i}\log\left(\sum_k \varphi_{kw}\widehat{\beta}_{ik}\right)}{|L_i|}\right), \quad (9)$$

where φ is the estimated topic-word distribution from the topic model under consideration.

Results. We compare the various benchmarks based on both truth measures, using perplexity score. For each benchmark, we compute the performance of the estimates based on each query from each participant in each category separately. We then compute the average performance over the queries submitted by each participant in each category. We report the average performance across all participants in Table 8,

Table 5. Topics Extracted from HDLDA

Topic	Examples of relevant words	Example of query	Example of web page
Ski			
1	Family, lesson, kids, rental, beginner	"Vermont family friendly ski resort"	Home page of the ski resort Mad River Mountain
2	Hotel, spa, reviews, luxury, exclusive	"Ski Vermont spa"	Exclusive ski package from Killington on tripadvisor
Printer			
1	Wireless, buy, shipping, black, scan	"Cheap printer copier scanner"	Brother wireless all-in-one printer on Amazon
2	Photo, ink, quality, pro, wifi	"Home lab quality printer"	PIXMA iP4000R photo printer on U.S.A. Canon
3	Printing, student, campus, double, duration	"Student printer"	On-campus student printing service info
Car			
1	Miles, mpg, price, dealer, fuel	"Used car Prius"	A used Honda Accord on Cargurus.com
2	Honda, Toyota, Nissan, Volkswagen, safety	"Big fuel efficient cars"	A list of luxury crossover SUVs on USNews
3	Electric, play, insurance, small, home	"best small city car"	Blog on whether to lease or buy a new car
Laptop			
1	Amazon, shipping, accessories, customer, buy	"Buy Windows laptop"	Best laptops of 2015 on CNET.com
2	Business, play, Lenovo, thinkpad, ideapad	"Lenovo y50"	Laptop reviews on lenovo.com
3	Battery, gaming, performance, display, dell	"Laptop long battery life"	Gaming laptop guide on tomsguide.com
4	Intel, CPU, core, ram, mainboard	"Intel core i5 CPU"	Intel core and AMD comparison on cpuboss.com
Camera			
1	Waterproof, kids, screen, touch, tablet	"kid friendly waterproof camera"	Polaroid waterproof digital camera on Kmart
2	Lens, DSLR, ISO, compact, shot	"Nikon d3200 bundle"	Nikon D5300 review on Camera Labs

where K for each category is set to be the same as that in Section 5.2 (i.e., the most interpretable set of topics). For robustness, we replicate Table 8 while setting K to be the same across product categories for $K \in \{2, 3, 4\}$ in Appendix D.[16]

Taking the average performance across categories gives us an average performance for each participant. In the last column of Table 8, we report the average performance across participants. We compare all the benchmarks using paired two-sample t-tests. We first consider the results when the truth measure is based on the chosen web page. We can see that HDLDA (strategic), which leverages the output of HDLDA and assumes that users leverage semantic relationships when forming queries, provides significantly better perplexity than the other benchmarks that assume users do not leverage these relationships. Both naive benchmarks, HDLDA (naive) and LDA, perform similarly to one another overall. When the truth measure is based on the task description, the comparisons are, in general, consistent with those using the other truth metric with the exception of the camera category. The average pattern also holds when setting different values of K (see Appendix D).

In conclusion, our results suggest that the output of HDLDA may be used as a basis for estimating content preferences from queries and that the assumption that users strategically leverage semantic relationships when formulating queries leads to estimates that are more accurate than those reached under the assumption that users naively formulate queries that directly reflect their content preferences.

6. Field Application

Our lab experiment provided us with some (imperfect) measure of consumers' actual content preferences. In this section, we illustrate the use of HDLDA in practice, using field data from a company that heavily relies on search advertising on Google. In particular, we explore whether the content preferences estimated from HDLDA may be used to explain and predict consumer click-through behavior. A sponsored search ad usually contains a heading, a link, and ad copy (a short description/preview of the landing page, shown to the user on the search engine results page). Figure 6 shows four examples of search ads that may appear on Google when searching for "vacation package Florida." One can see that the search ads shown in response to a given search query may contain very different headings and descriptions. One key performance metric of a search

Table 6. Mean and Standard Deviation of θ_q and θ_p Within Each Category

Category	Parameter	Topic 1	Topic 2	Topic 3	Topic 4
Ski	θ_q	0.55 (0.50)	0.45 (0.50)		
	θ_p	0.62 (0.31)	0.38 (0.31)		
Printer	θ_q	0.46 (0.41)	0.35 (0.39)	0.18 (0.29)	
	θ_p	0.33 (0.30)	0.41 (0.30)	0.26 (0.28)	
Car	θ_q	0.20 (0.31)	0.45 (0.40)	0.35 (0.39)	
	θ_p	0.19 (0.19)	0.52 (0.28)	0.30 (0.29)	
Laptop	θ_q	0.21 (0.28)	0.17 (0.24)	0.36 (0.33)	0.25 (0.30)
	θ_p	0.30 (0.27)	0.19 (0.20)	0.43 (0.27)	0.08 (0.13)
Camera	θ_q	0.43 (0.49)	0.57 (0.49)		
	θ_p	0.49 (0.32)	0.51 (0.32)		

Note. Standard deviations are in parentheses.

Table 7. Posterior Estimates of the Topic Intensities of Task Descriptions

	Task	Topic 1	Topic 2	Topic 3	Topic 4
Ski	1	0.85	0.14		
	2	0.69	0.31		
Printer	3	0.14	0.66	0.20	
	4	0.01	0.98	0.01	
Car	5	0.06	0.16	0.78	
	6	0.06	0.02	0.92	
Laptop	7	0.02	0.03	0.90	0.05
	8	0.01	0.30	0.68	0.01
Camera	9	0.99	0.01		
	10	0.22	0.78		

ad campaign is CTR. It is well known by practitioners and academics that position influences CTR significantly (Kihlstrom and Riordan 1984, Hotchkiss et al. 2005, Varian 2007, Agarwal et al. 2011). However, there has been very little academic research investigating the impact of the *copy* of an online ad on CTR.

All else equal, CTR should be higher for a sponsored search ad whose copy is better aligned with the content preferences of consumers who type the corresponding query. If this is the case, the degree of alignment between content preferences estimated based on HDLDA and the copy of the ad should be predictive of CTR. We test whether this is the case, using sponsored search data from an advertiser on Google. First, we estimate HDLDA from a subset of the queries on which the firm advertises on Google and the corresponding organic search results. Then, following the procedure given in Section 3.3, we use the output from HDLDA to estimate content preferences underlying each search query on which the firm advertises and the topic intensities of each ad copy used by the firm. Finally, we test whether the CTR for a (search query, ad copy) pair is linked to the degree to which the topic intensities of the ad copy shown on the search engine results page match with the content preferences estimated based on the query, controlling for various factors, such as quality score and position.

6.1. Data

Our data came from a large global online portal, on which consumers can book airline tickets, hotel rooms, and rental cars. We only consider search queries that were matched based on either "exact" or "phrase" keyword match.[17] We focus on queries that are more relevant for the advertiser by only including queries that received at least eight impressions over the entire time window. Each observation in our data set concerns a combination of one search query and one ad copy. For each observation, we have access to the following information based on the firm's campaigns running from 2013 to 2016: total number of impressions, total number of clicks, text of the ad copy shown on the search engine results page, average position of the ad, and the quality score assigned by Google.[18] Our final data set consists of 13,069 (search query, ad copy) pairs with 12,856 unique search queries and 633 unique ad copies. We find that 98.65% of queries are matched to only one ad copy, and on average, each ad copy is matched to 20.65 queries with a standard deviation of 147.49. Table 9 provides summary statistics of all the variables.

6.2. HDLDA

To explore the ability of content preferences estimated based on HDLDA to predict CTR out of sample, we randomly select 3,000 unique queries from our data as training queries for HDLDA and set the others aside for out-of-sample validation. We again use Google customer search API to collect the top 10 organic search results for these queries in the training data and use a Python script to download the web page content of all the associated organic search results. This results in 6,578 unique URLs. We process all the textual

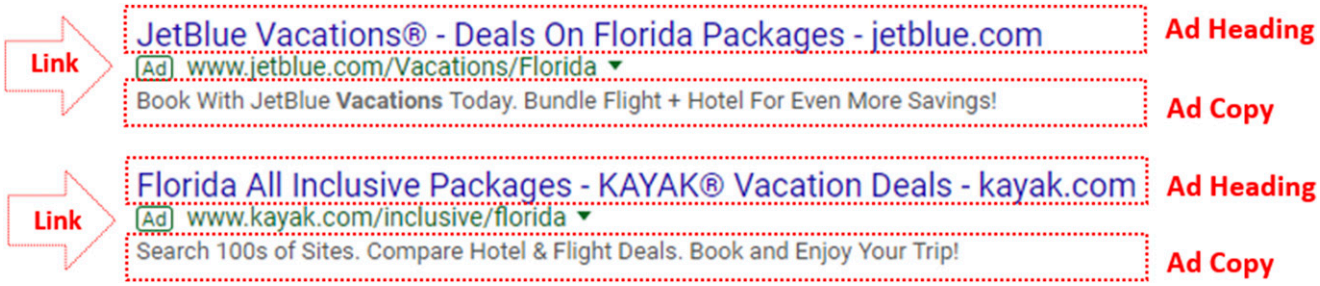
Table 8. Estimating Content Preferences from Queries

	Ski	Printer	Car	Laptop	Camera	
	$K = 2$	$K = 3$	$K = 3$	$K = 4$	$K = 2$	Average
Chosen web page						
HDLDA (strategic)	187*	364	378*	453*	351*	346*
HDLDA (naive)	205	366	508	704	589	475
LDA	202	368	461	658	551	448
Task description						
HDLDA (strategic)	275*	167	394*	614*	294	348*
HDLDA (naive)	318	151*	420	712	267*	373
LDA	308	161	409	700	270*	369

Notes. We compute the *perplexity* score of the "true" description of each participant's content preferences (i.e., task description or chosen web page) given the estimated content preferences. Smaller perplexity indicates better performance. The last column is the average of the average performance for each participant across all the tasks.

*Model is best or tied for best at $p < 0.05$.

Figure 6. (Color online) Search Ad Example



information in the search queries and the web page content, following the procedure described in Section 4.4. The vocabulary consists of 11,421 unique terms. Given that the vocabulary is much larger in this field study compared with the lab experiment, the number of topics required for HDLDA is also higher. We select the number of topics K based on trading off fit, interpretability, and computational considerations and set $K = 20$. Because all the sampling is independent across web pages and across queries, we use parallel computing to speed up the estimation.[19]

6.3. Regression Analysis

Given the estimated topics φ from HDLDA, we then estimate the topic intensities of all the search queries (including the out-of-sample queries) and ad copies. Next, we estimate content preferences for each query based on HDLDA (strategic), HDLDA (naive), and LDA, using the approach described in Section 3.3. Finally, we compute the similarity between the estimated topic intensities of a given ad copy a, $\widehat{\theta}_a$, and the estimated consumer content preferences behind query q, $\widehat{\beta}_q$, using cosine similarity: $\cos(\widehat{\theta}_a, \widehat{\beta}_q)$. This variable measures the extent to which a given ad copy matches with the estimate of content preferences based on the corresponding query.

We use a logit function to link CTR for the (search query, ad copy) pair (q,a) to the independent variables:

$$\mathrm{CTR}_{qa} = \frac{\exp(U_{qa})}{1+\exp(U_{qa})}, \tag{10}$$

where

$$U_{qa} = \delta_a + \mu_1 \mathrm{Cos}(\widehat{\theta}_a, \widehat{\beta}_q) + \mu_2 \mathrm{Position}_{qa} + \mu_3 \mathrm{AdQuality}_{qa} + \mu_4 \mathrm{Length}_q + \mu_5 \mathrm{Length}_a \tag{11}$$

and δ_a denotes random effects for ad copy. That is, we study whether the similarity between the estimated content preferences and the topic intensities of the ad copy shown on the search engine results page predicts CTR, controlling for position effects, ad quality score, the lengths of the query and the ad copy, and ad copy random effects. Only cosine similarity differs across benchmarks.

We estimate this regression model with the cosine similarity computed from each of the three approaches, using maximum likelihood. We also estimate this model without cosine similarity, which we label as "No Content." Table 10 reports the results. We find that the cosine similarity between the content of the ad copy and the content preferences estimated based on the query using HDLDA (strategic) is significantly positively related to CTR ($p < 0.05$) even when controlling for ad copy random effects, quality score, position, and other covariates. In contrast, when content preferences are estimated based on HDLDA (naive) or LDA, cosine similarity is not significant. The effect of the other variables is as expected with CTR significantly decreasing with position and increasing with quality score.

To test predictive validity, we reestimate these regression models only based on the 3,000 queries that were used to train HDLDA and use the regression

Table 9. Summary Statistics of Field Data

Variable	Mean	Standard deviation	Minimum	Maximum
Number of impressions	454	10,621	8	726,943
Number of clicks	131	4999	1	500,190
CTR	0.365	0.214	0.002	1
Average position	1.824	0.824	1	12.200
Ad quality score	8.914	0.833	5	10
Length of query	3.927	1.389	1	22
Length of ad copy	13.584	0.621	12	16

Notes. The unit of observation is a (search query, ad copy) pair. Our data set contains 13,069 such observations with 12,856 unique search queries and 633 unique ad copies. The length of a query or an ad copy is computed based on the original number of terms, not the words in the vocabulary.

Table 10. Field Study: Regression Results

	HDLDA (strategic)	HDLDA (naive)	LDA	No content
Copy random effect	–0.642 (0.176)	0.104 (0.185)	0.096 (0.186)	0.113 (0.187)
Position	–0.500***	–0.499***	–0.498***	–0.500***
Ad quality score	0.082***	0.085***	0.085***	0.089***
Length of query	0.016	0.019	0.019	0.019
Length of ad copy	–0.032	–0.053	–0.053	–0.053
Cosine similarity	0.817**	0.069	0.070	
Number of observations	13,069	13,069	13,069	13,069
AIC	2,314	2,319	2,319	2,319
MAE (in-sample)	0.141	0.143	0.145	0.143
MAE (out-of-sample)	0.155	0.157	0.159	0.158

Notes. For copy random effect, mean is reported with standard deviation in parentheses. Based on paired two-sample *t*-tests, both the out-of-sample and in-sample MAE from HDLDA (strategic) are significantly smaller than the other models ($p < 0.05$).

***$p < 0.01$; **$p < 0.05$; *$p < 0.01$ (for regression estimates).

estimates to predict consumer CTR on the remaining queries, which were not used to train HDLDA or the regression model. We report the mean absolute error (MAE) as our metric for prediction accuracy in Table 10. For in-sample prediction accuracy, we find that the absolute error from HDLDA (strategic) is significantly lower than that of the other three models ($p < 0.05$). For out-of-sample prediction, HDLDA (strategic) is significantly better than any of the other three models ($p < 0.01$). Although the improvement in prediction is relatively modest, given that online search advertising is a $90 billion industry (Statista 2017), it still has the potential to have significant financial impact.

To sum up, the field study illustrated one potential practical application of HDLDA. We find that content preferences estimated based on HDLDA may be used by firms to improve their predictions of CTR for sponsored ads even after controlling for traditional predictors, such as position and quality score. Such predictive ability is critical to improve the effectiveness of SEM campaigns. For instance, the cosine similarity computed from HDLDA could be useful in identifying more relevant ad copy for a given search query without spending time and resources testing experimentally each potential (ad copy, search query) pair. The firm with which we collaborated on this study is currently exploring using our approach to improve the selection of ad copies from thousands of available designs for thousands of their target queries.

7. Discussion and Conclusion

In this paper, we develop a new topic model, HDLDA, that jointly estimates the topic intensities in queries and web pages as well as the mapping between queries and their results. In our domain of application, HDLDA captures the facts that a web page is retrieved by certain queries and that topics in queries are semantically related to topics in search results. More generally, HDLDA is a model for bag-of-word data that can be applied to any context in which one type of documents are semantically related to another type of documents. HDLDA has a new structure within the broad literature of topic modeling, and our paper provides a methodological contribution to the probabilistic topic-modeling literature.

Using the output of HDLDA, it is possible to estimate a consumer's content preferences on the fly based on each query. For example, if we make the assumption that consumers strategically formulate queries that will retrieve content that matches their preferences in expectation, we can estimate a consumer's content preferences as the expected topic intensities in the search results given the topic intensities in the search query. Alternatively, if we make the assumption that consumers naively formulate search queries that directly reflect their preferences, content preferences may be estimated as the expected topic intensities in the search query itself. Our data suggest that content preferences estimated based on HDLDA and assuming that consumers are strategic are more accurate than content preferences estimated under the naive assumption based on either HDLDA or a standard LDA model.

From a managerial perspective, HDLDA can automatically extract, understand, and organize the meaning of queries and web pages within a search domain without human intervention. We illustrate one practical managerial application of HDLDA to the prediction of CTR in sponsored search advertising. As another illustration, consider a tech product review website, such as CNET.com, which produces content related to laptops, one of the categories featured in our lab experiment. The website could use HDLDA to estimate the content preferences associated with any query and then compute the fit (measured by the cosine similarity) between these preferences and different web pages. For example, suppose CNET.com wanted to promote a web page about getting a Lenovo Y50 touch gaming laptop.[20]

The firm would be able to infer, for instance, that the query "student personal laptops" is more relevant to this web page compared with the query "durable lightweight laptop" (cosine similarity of 0.96 versus 0.52 based on the estimates from our lab study and the actual content of the web page). In the context of SEO, such information could help the website decide that it should attempt to improve the organic search ranking of this web page for the more relevant query. In the context of SEM, this information would help the firm decide to have that web page appear as a sponsored ad for the more relevant query. More generally, the output of HDLDA can guide firms' SEO and SEM strategies by helping them quantify how well their content (web page or ad copy) matches the content preferences captured by various queries and focus their efforts on promoting their content for those queries with better fit in an efficient and interpretable manner.

We close by highlighting additional areas for future research. First, to validate our approach for estimating content preferences, we developed our own search engine. By allowing the removal of variations in organic and sponsored search results from customization, this tool offers opportunities to shed new light on important research questions in consumer search and search advertising, such as position effects and advertising effects. Indeed, although we did not do this in the current paper, this tool allows varying the order of organic and sponsored search results, moving organic results to the sponsored section, etc. Second, future research might explore alternative assumptions on the way consumers translate their preferences into search queries and on their beliefs and knowledge of the semantic relationships between queries and results. Finally, future research may combine information on queries with information on clickstream behavior to provide a more extensive set of observations based on which content preferences may be estimated. Recent developments in collaborative topic modeling (Wang and Blei 2011) might provide the foundation for models that would formulate both query formation and clicking behavior as functions of content preferences.

Acknowledgments
This paper is based on the second chapter of Jia Liu's doctoral dissertation.

Appendix A. Inference Algorithm for HDLDA
This appendix describes the inference algorithm for HDLDA. Across the collection of Q queries, let $n^k_{q,j}$ be the number of word tokens in the qth query that are the jth word in the vocabulary and are assigned to the kth topic. Mathematically, $n^k_{q,j} = \sum_{i=1}^{J_q} I(z_{qi} = k \wedge w_{qi} = j)$. Hence, $n^k_{q,j}$ is three-dimensional: query, word, and topic. By summing the counts across different words within a query, we get N^k_q, which denotes the number of words that are assigned to topic k in query q. By summing the counts across different queries, we get N^k_j, which denotes the number of queries that have the jth word in the vocabulary and are also assigned to the kth topic. Across the collection of P web pages, we defined $m^k_{p,j}$ as the number of word tokens in the pth web page that are the jth word in the vocabulary and that are assigned to the kth topic; that is, $m^k_{p,j} = \sum_{i=1}^{J_p} I(z_{pi} = k \wedge w_{pi} = j)$. We similarly define the summation across words and web pages, respectively, which are denoted as M^k_p and M^k_j.

Gibbs Sampler for Assignments z_p and z_q
Given the topic intensities θ_q and the word distribution φ, the posterior distribution of each z_{qj} is

$$\Pr(z_{qj} = k \mid w_{qj}, \theta_q, \{\varphi_k\}) = \frac{\Pr(w_{qj} \mid z_{qj} = k, \varphi_k)\Pr(z_{qj} = k \mid \theta_q)}{\sum_{i=1}^{K}\Pr(w_{qj} \mid z_{qj} = i, \varphi_i)\Pr(z_{qj} = i \mid \theta_q)} = \frac{\varphi_{k,w_{qj}}\theta_{qk}}{\sum_{i=1}^{K}\varphi_{i,w_{qj}}\theta_{qi}}. \tag{A.1}$$

Similarly, the posterior distribution of the assignment z_{pi} depends on the data w_{pi} and the latent distributions φ and θ_p. The posterior distribution of each z_{pi} is

$$\Pr(z_{pi} = k \mid w_{pi}, \theta_p, \{\varphi_k\}) = \frac{\varphi_{k,w_{pi}}\theta_{pk}}{\sum_{j=1}^{K}\varphi_{j,w_{pi}}\theta_{pj}}. \tag{A.2}$$

Gibbs Sampler for φ and θ_p
The posterior of the topic distribution φ in the collection is still a Dirichlet and only depends on the latent assignment and the data including both queries and web pages:

$$\Pr(\varphi_k \mid \{z_p\}, \{z_q\}, \{w_q\}, \{w_p\}, \eta) = Dir_J(\eta + N^k_1 + M^k_1, \ldots, \eta + N^k_J + M^k_J). \tag{A.3}$$

The posterior distribution of the topic intensities θ_p for each web page p only depends on its latent assignment and its prior. This distribution is also given in closed form, conditional on the semantic relationship matrix R and $\{\theta_q\}$:

$$\Pr(\theta_p \mid z_p, R, \{\theta_q\}, l_p) = Dir_K(\exp(R^T_1\overline{\theta}_q(p)) + M^1_p, \ldots, \exp(R^T_K\overline{\theta}_q(p)) + M^K_p). \tag{A.4}$$

Metropolis–Hastings Algorithm for θ_q
The posterior distribution of the topic intensities θ_q for each query q is nonconjugate. Indeed, it depends not only on the latent assignment z_q, but also on the topic intensities of the web pages that can be retrieved by query q. We use Metropolis-within-Gibbs and apply the iteration sequentially to each q:

$$\begin{aligned}&\Pr(\theta_q \mid z_q, \{\theta_p\}, \{\theta_{-q}\}, R, \{l_{pq}\}, \alpha)\\ &\quad\propto \Pr(\{\theta_p\} \mid \theta_q, \theta_{-q}, R, \{l_{pq}\})\Pr(z_q \mid \theta_q)\Pr(\theta_q \mid \alpha)\\ &\quad\propto \prod_{p, l_{pq}=1} Dir_K\left(\theta_p \mid \left\{\exp\left(R^T_k\frac{\sum_{q\in Q}\theta_q l_{pq}}{\sum_{q\in Q} l_{pq}}\right)\right\}\right)\\ &\quad\cdot Dir_K(\theta_q \mid \alpha + N^1_q, \ldots, \alpha + N^K_q).\end{aligned} \tag{A.5}$$

Here, we apply an adaptive proposal distribution (with vanishing adaptation) for a random walk Metropolis–Hastings algorithm to sample each θ_q (Andrieu and Thoms 2008). The proposal distribution is Dirichlet, $\theta_q^{(t)} \sim Dir(\sigma_{q,t}\theta_q^{(t-1)})$. We adaptively choose $\sigma_{q,t}$ for each θ_q to attain a target acceptance rate while preserving the convergence of the Markov chain by the Robbin–Monro algorithm:

$$\sigma_{q,t} = \sigma_{q,t-1} \exp\left((a^* - a_{q,t-1})/t^{\delta}\right),$$

where $a_{q,t-1}$ is the acceptance rate at iteration $t-1$ for θ_q; a^* is the optimal acceptance rate, which usually is set to 0.23 for large problems (Gelman and Meng 1996); and $\delta \in (0,1]$ controls the decay rate of the adaption. Basically, the precision will increase if the current acceptance rate is below the target rate. Note that, because the proposal distribution is asymmetric, the acceptance ratio for θ_q at each iteration t should be obtained as

$$r_{qt} = \min\left\{\frac{L(\theta_q^{(t)}) f(\theta_q^{(t-1)} \mid \sigma_{q,t}\theta_q^{(t)})}{L(\theta_q^{(t-1)}) f(\theta_q^{(t)} \mid \sigma_{q,t}\theta_q^{(t-1)})}, 1\right\},$$

where $f(x|y)$ denotes the density of the Dirichlet distribution $Dir(y)$ at x. Based on our empirical study, we find that, for a small corpus, even a random walk Metropolis–Hastings (without adaptation) algorithm converges quite well for sampling θ_q.

Maximizing R from a Dirichlet-Multinomial Regression Model

The parameter R controls the relationship between θ_q and θ_p, which is captured by a Dirichlet-multinomial regression model in Equation (3). We estimate R by optimizing the full log-likelihood of the model (Mimno and McCallum 2008):

$$l(R) = \sum_{p=1}^{P}\left(\log\Gamma\left(\sum_k \exp(R_k^T \overline{\theta}_q(p))\right) - \sum_k [\log\Gamma(\exp(R_k^T \overline{\theta}_q(p))) - (\exp(R_k^T \overline{\theta}_q(p)) - 1)\log(\theta_{pk})]\right).$$

The derivative of this log-likelihood with respect to r_{tk} is

$$\frac{\partial l(R)}{\partial r_{tk}} = \sum_{p=1}^{P} \overline{\theta}_{qt}(p) \exp(R_k^T \overline{\theta}_q(p))\left(\Psi\left(\sum_j \exp(R_j^T \overline{\theta}_q(p))\right) - \Psi\left(\exp(R_k^T \overline{\theta}_q(p))\right) + \log(\theta_{pk})\right),$$

where $\Psi(\cdot)$ denotes the digamma function that is defined as the logarithmic derivative of the gamma function, $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$. The optimization problem could be difficult if K is large. Our implementation is mainly based on the standard Broyden–Fletcher–Goldfarb–Shanno optimization because this method has been shown to be fast, robust, and reliable in practice.

(Optional) Maximizing η from Its Likelihood Function

The symmetric prior parameter η controls the prior of φ. Similar to before, we estimate η by optimizing the joint-likelihood of $\varphi_v \sim Dirichlet_K(\eta)$ for $v = 1, 2, \ldots, J$.

Appendix B. Simulation Study

This appendix presents a synthetic data analysis of HDLDA. We first describe the data-generation process of the model and the parameterization of our simulation study. Then we summarize the estimation results based on the inference algorithm presented in Appendix A.

Data Generation

For a given set of parameters $\{K, Q, P, J, \{J_q\}_q, \{J_p\}_p, \{l_{pq}\}_{p,q}, \eta, \alpha, R\}$, the following procedure describes the data-generative process for HDLDA:

1. For each topic $k = 1, 2, \ldots, K$, draw a distribution over words: $\varphi_k \mid \eta \sim Dirichlet_J(\eta)$
2. For each query $q = 1, 2, \ldots, Q$,
   (a) Draw topic intensities $\theta_q \mid \alpha \sim Dirichlet_K(\alpha)$
   (b) For $j = 1, 2, \ldots, J_q$,
      i. Draw topic assignment $z_{qj} \mid \theta_q \sim Category(\theta_q)$
      ii. Draw word $w_{qj} \mid (z_{qj}, \{\varphi_k\}) \sim Category(\varphi_{z_{qj}})$
3. For each web page $p = 1, 2, \ldots, P$,
   (a) Calculate the average topic intensities among its labeling queries: $\overline{\theta}_q(p) = \frac{\sum_q \theta_q l_{pq}}{\sum_q l_{pq}}$
   (b) Draw topic intensities $\theta_p \mid (R, \overline{\theta}_q(p)) \sim Dirichlet_K(\exp(R_1^T \overline{\theta}_q(p)), \ldots, \exp(R_K^T \overline{\theta}_q(p)))$
   (c) For $i = 1, 2, \ldots, J_p$,
      i. Draw topic assignment $z_{pi} \mid \theta_p \sim Category(\theta_p)$
      ii. Draw word $w_{pi} \mid (z_{pi}, \{\varphi_k\}) \sim Category(\varphi_{z_{pi}})$

We simulate a data set with a structure similar to real search data that we collected from the experimental study described in Section 4. Specifically, we set $K = 3$, $Q = 800$, $P = 4{,}000$, and $J = 2{,}000$. For $q = 1, 2, \ldots, Q$, we draw the integer J_q randomly (uniformly) from [2,20]; for $p = 1, 2, \ldots, P$, we draw the integer J_p randomly (uniformly) from [300,600]. We draw $\{l_{pq}\}$ so that each query can retrieve 10 pages, and each page can be retrieved by at least one query (the mean is 2.00 with a standard deviation 1.02). For the mapping matrix R, we set all the diagonal elements to be 0.8 and set all the off-diagonal elements to be 0.4. We set $\alpha = 1$ and $\eta = 0.01$. All other parameters are generated according to the process described previously.

We calibrate HDLDA on this simulated data set using the inference algorithm described in Appendix A. The model parameters include about 1.79 million latent word assignments z, 6,000 parameters in $\{\varphi_k\}$, 12,000 parameters in $\{\theta_p\}$, 2,400 parameters in $\{\theta_q\}$, and nine parameters in R. We run 10,000 MCMC iterations and use the first 5,000 as burn-in.

Simulation Results

Before presenting the estimation results, we provide some background on the identification of topic models (especially LDA) in general. This is important in forming reasonable expectations on what and how much can be recovered in HDLDA. Although topic modeling is an approach that has proved successful in automatic comprehension and classification of data, only recent work has attempted to give provable guarantees for the problem of learning the model parameters (Anandkumar et al. 2012, Arora et al. 2012). The problem of recovering nonnegative matrices φ (topics) and θ (topic intensities) with small inner-dimension K (number of

Table B.1. Simulation Results for $\{\varphi_k, \theta_p, \theta_q\}$

Parameters	Coverage	Mean squared error
$\{\varphi_k\}$	90.12%	1.06e-09
$\{\theta_p\}$	89.77%	0.0006
$\{\theta_q\}$	74.67%	0.0563

topics), is NP hard (Arora et al. 2012). As a solution, recent work has relied on very strong assumptions about the corpus, for example, restricting one topic per document or assuming each topic has words that appear only in that topic. At best, φ could only be recovered up to permutations (Anandkumar et al. 2012). In addition, according to Arora et al. (2012), it is impossible to learn the topic intensities matrix θ to within arbitrary accuracy, and this is theoretically impossible even if we knew φ and the distribution from which θ is generated.

Given this background, empirically one should not expect all parameters in topic models to be recovered. As an example, we test the well-known Gibbs sampler algorithm on a basic LDA, using multiple simulated corpora that have similar size as the one described herein. We find that about 90% of the topics φ are covered by the 95% credible interval (CI), and about 80% of the topic proportions θ are covered by the 95% CI. Given the complexity of the inference algorithm for HDLDA, one should not expect its recovery to be better than the Gibbs sampler for a LDA.

As measures of recovery performance, we report the proportion of the parameters that are recovered by the posterior 95% CI and the mean square error (MSE) between the true and the estimated parameters. The details are given in Table B.1 for $\{\varphi_k\}$, $\{\theta_p\}$, and $\{\theta_q\}$, respectively. One can see that the recovery for both $\{\varphi_k\}$ and $\{\theta_p\}$ is pretty good, and it is decent for $\{\theta_q\}$. Finally, we report the posterior estimates and the 95% CI for all the parameters in R, which are given in Table B.2. Note that, because the maximum likelihood estimation (MLE) for the Dirichlet-multinomial regression has significant bias if the sample size is not large enough,[21] we would not expect the estimators of R from the stochastic EM algorithm to do better than the MLE of R using the true data (Nielsen 2000). That is, the best that the inference algorithm can achieve in estimating R is recovering its MLE, denoted as R_{MLE}, that is estimated using the true $\{\theta_p\}$ and $\{\theta_q\}$ and also reported in Table B.2. We can see for the true R, six out of nine parameters are covered by the 95% CI. This is increased to eight in recovering R_{MLE}.

Table B.2. Simulation Results for R

Parameters	R	R_{MLE}	Posterior estimate	95% CI
r_{11}	0.8	0.841	0.622	[0.547, 0.723]
r_{21}	0.4	0.337	0.410	[0.307, 0.487]
r_{31}	0.4	0.369	0.442	[0.349, 0.537]
r_{12}	0.4	0.439	0.403	[0.339, 0.463]
r_{22}	0.8	0.810	0.780	[0.685, 0.839]
r_{32}	0.4	0.338	0.351	[0.293, 0.426]
r_{13}	0.4	0.413	0.365	[0.262, 0.389]
r_{23}	0.4	0.429	0.583	[0.510, 0.672]
r_{33}	0.8	0.722	0.698	[0.584, 0.798]

Appendix C. Lab Study—Statistics About Users' Chosen Web Pages

Table C.1 provides the average cosine similarity in the topic intensities between the chosen web page and any clicked but nonchosen web page within each category and under each K for each benchmark. Similarly, Table C.2 provides the average cosine similarity between the chosen web page and all search results shown by the search engine to the participant in response to that query.

Table C.1. Average Cosine Similarity in Topic Intensities Between Chosen Web Page and Clicked (Nonchosen) Web Pages

	$K = 2$		$K = 3$		$K = 4$	
	HDLDA	LDA	HDLDA	LDA	HDLDA	LDA
Ski	0.899	0.859	0.877	0.870	0.771	0.766
Printer	0.842	0.812	0.732	0.745	0.666	0.592
Car	0.875	0.853	0.798	0.818	0.719	0.701
Laptop	0.955	0.962	0.802	0.735	0.705	0.631
Camera	0.827	0.778	0.745	0.689	0.608	0.517

Table C.2. Average Cosine Similarity in Topic Intensities Between Chosen Web Page and All Search Results

	$K = 2$		$K = 3$		$K = 4$	
	HDLDA	LDA	HDLDA	LDA	HDLDA	LDA
Ski	0.875	0.837	0.842	0.829	0.721	0.723
Printer	0.784	0.750	0.697	0.675	0.588	0.501
Car	0.823	0.789	0.744	0.750	0.651	0.638
Laptop	0.950	0.959	0.778	0.695	0.691	0.612
Camera	0.801	0.746	0.716	0.656	0.583	0.485

Appendix D. Lab Study—Model Evaluation for $K \in \{2, 3, 4\}$

We replicate the analysis in Table 8 while setting K to be the same across all the product categories for $K \in \{2, 3, 4\}$. Smaller perplexity indicates better performance. The asterisk means that a model is best or tied for best at $p < 0.05$. Results are presented in Tables D.1 ($K = 2$), D.2 ($K = 3$), and D.3 ($K = 4$).

Table D.1. Estimating Content Preferences from Queries: $K = 2$

Model	Ski	Printer	Car	Laptop	Camera	Average
Chosen web page						
HDLDA (strategic)	187*	362*	382*	448*	351*	346*
HDLDA (naive)	205	451	476	876	589	519
LDA	202	416	430	800	551	480
Task description						
HDLDA (strategic)	275*	212*	387*	627*	294	359*
HDLDA (naive)	318	218	424	997	267*	445
LDA	308	233	412	918	270*	428

*Model is best or tied for best at $p < 0.05$.

Table D.2. Estimating Content Preferences from Queries: $K = 3$

Model	Ski	Printer	Car	Laptop	Camera	Average
Chosen web page						
HDLDA (strategic)	193	364	378*	449*	357*	348*
HDLDA (naive)	159*	366	509	776	563	475
LDA	159*	368	461	710	476	435
Task description						
HDLDA (strategic)	112	168	394*	620*	299	318*
HDLDA (naive)	105*	151*	420	765	260	340
LDA	102*	161	409	749	251*	334

*Model is best or tied for best at $p < 0.05$.

Table D.3. Estimating Content Preferences from Queries: $K = 4$

Model	Ski	Printer	Car	Laptop	Camera	Average
Chosen web page						
HDLDA (strategic)	191	367*	376*	454*	346*	347*
HDLDA (naive)	173	478	516	704	356	445
LDA	166*	440	459	658	492	443
Task description						
HDLDA (strategic)	115*	171	399*	614*	296	319*
HDLDA (naive)	125	152*	457	712	290*	347
LDA	113*	157	424	700	294*	337

*Model is best or tied for best at $p < 0.05$.

Endnotes

[1] We acknowledge that topic proportions only refer to the distribution of information within a document, not the absolute volume of information. Some pages might have a topic distribution that is further away from the consumer's "ideal point" and yet preferred because there is just more information overall on this page.

[2] One may suggest to use the summation over all the θ_q's rather than their average in this regression model. However, this would artificially decrease the variance of the Dirichlet distribution for web pages that are retrieved by more queries.

[3] Although one can also estimate these parameters using an additional Metropolis–Hastings step, Mimno and McCallum (2008) have suggested that a stochastic EM sampling approach works very efficiently for topic models. Because the R matrix is estimated by maximization rather than simulation, we admit that this may underestimate the variance of the elements in R.

[4] Another option would be to estimate a different symmetric parameter η_k for each topic.

[5] Note that in Equations (5)–(7), β_i is estimated without observing the actual search results corresponding to the query.

[6] These were relevant categories for our participants who are mostly undergraduate and graduate students. We ran the lab experiment in the winter on the east coast of the United States, so ski resorts were also relevant.

[7] Note that we designed these task descriptions before acquiring and analyzing any web page content; that is, our task descriptions were *not* designed to be aligned with or map onto the set of topics from HDLDA. Hence, we should expect each task description to have positive intensities on multiple topics.

[8] In our case, the edit distance counts the minimum possible weighted number of character operations (including insertions, deletions, and substitutions) required to transform one query into the other. For example, the edit distance between "best school laptop" and "school laptop" is five (five characters need to be deleted: b, e, s, t, space).

[9] We first select words that appear at least n times in total (n is between 10 and 20 depending on the corpus based on inspection—the selection of n was finalized before running HDLDA). This helps eliminate a fair amount of meaningless words that cannot be removed in preprocessing. Then, we calculate the mean term frequency-inverse document frequency ($tf\text{-}idf$) for the remaining words. The $tf\text{-}idf$ is commonly used to select vocabulary in topic modeling and is computed as $tf\text{-}idf(w) = f_w \times \log\left(\frac{N}{n_w}\right)$, where f_w is the average number of times word w occurs in each document in the corpus and n_w is the number of documents containing word w. We keep words whose $tf\text{-}idf$ is above the median, which allows omitting words that have low frequency as well as those occurring in too many documents (Hornik and Grün 2011).

[10] In some cases, the total number of links could be smaller than 10 for a query if the query either cannot be understood by Google or has very few relevant results or if scripting the content of a web page is prohibited by the website.

[11] We also tried other potential evaluation metrics for Bayesian models, such as DIC and Watanabe–Akaike information criterion (Gelman et al. 2014). However, they all favor unreasonably large K.

[12] For each corpus and a given K, we run 6,000 MCMC iterations using the fist 4,000 as burn-in, saving every fifth iteration. We evaluate the convergence of the MCMC sequence by plotting the time series and conducting Geweke convergence diagnostics (Geweke 1991).

[13] In the data, we find some participants did not follow the instructions correctly (e.g., participants submitted some random text without actually conducting the search on Hoogle or found their chosen link on other search engines). We drop these observations in our analysis. The proportion of observations dropped is 2% for the ski category, 2% for printer, 10% for car, 12% for laptop, and 6% for camera. As a result, there are 195 participants who submitted a valid chosen web page for at least one task.

[14] We estimate the topic intensities of the task descriptions using the same procedure as we use to estimate $\widehat{\theta}_q$ based on the output of HDLDA. We report the posterior mean over 2,000 MCMC iterations.

[15] The cosine similarity between two vectors a and b is $f(a,b) = \frac{a \cdot b}{||a||||b||} = \frac{\sum_k a_k b_k}{\sqrt{\sum_k a_k^2}\sqrt{\sum_k b_k^2}}$.

[16] We also replicate all the tables in Appendix D while fixing the hyperparameter for the topic distribution of queries α to be 0.01 or one (rather than 0.1). We find that the overall pattern of results is robust to these different values of α. Results are available from the authors.

[17] An exact or a phrase match ensures that the actual search query typed by consumers contains the keywords in the same order.

[18] We only have the quality score at the time when the company pulled this data set for us, not its entire history. This should not reduce the predictive validity of quality score relative to $\cos(\widehat{\theta}_a, \widehat{\beta}_q)$ (described in Section 6.3) as the latter is also based on data collected around the same time.

[19] The R matrix contains K^2 unknown parameters. It becomes computationally costly when K is more than 20 as the algorithm needs to solve an optimization problem with hundreds of parameters for every MCMC iteration. We used R programming language to run the estimation on a server with an Intel® Xeon® E7 processor using four physical cores. Each MCMC iteration takes around 60 seconds, depending on the choice of K. We run 5,000 iterations, using the first 3,000 as burn-in, and save every fifth iteration.

[20] The URL of this web page is http://www.cnet.com/news/get-a-lenovo-y50-touch-4k-gaming-laptop-for-999-99.

[21] For example, when we increase the number of web pages from 4,000 to 40,000, the MLE of R for the Dirichlet-multinomial regression model using the true $\{\theta_p\}$ and $\{\theta_q\}$ has much smaller bias. Specifically, the MSE is decreased from 0.0111 to 0.0006.

References

Abhishek V, Gong J, Li B (2018) Examining the impact of contextual ambiguity on search advertising keyword performance: A topic model approach. *MIS Quart*. Forthcoming.

Agarwal A, Hosanagar K, Smith MD (2011) Location, location, location: An analysis of profitability of position in online advertising markets. *J. Marketing Res.* 48(6):1057–1073.

Amerland D (2013) *Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact, and Amplify Your Online Presence*, 1st ed. (Que Publishing Company, Indianapolis, IN).

Anandkumar A, Foster DP, Hsu DJ, and Kakade SM, Liu Y-K (2012) A spectral algorithm for latent Dirichlet allocation. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Proc. 26th Annual Conf. Neural Information Processing Systems*, Vol. 1 (Curran Associates, Red Hook, NY), 917–925.

Andrieu C, Thoms J (2008) A tutorial on adaptive MCMC. *Statist. Comput.* 18(4):343–373.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Arora S, Ge R, Moitra A (2012) Learning topic models—Going beyond SVD. *IEEE 53rd Annual Sympos. Foundations of Computer Science* (IEEE, Piscataway, NJ), 1–10.

Bischof J, Airoldi EM (2012) Summarizing topical content with word frequency and exclusivity. *Proc. 29th Internat. Conf. Machine Learn*. (ACM, New York), 201–208.

Blei DM, Jordan MI, Griffiths TL, Tenenbaum JB (2003) Hierarchical topic models and the nested Chinese restaurant process. Thrun S, Saul LK, Scholkopf, B, eds. *Proc. 16th Annual Conf. Neural Information Processing Systems* (MIT Press, Cambridge, MA), 17–24.

Blei DM, Lafferty JD (2007) A correlated topic model of science. *Ann. Appl. Statist.* 1(1):17–35.

Blei DM, Lafferty JD (2009) Topic models. Srivastava AN, Sahami M, eds. *Text Mining: Classification, Clustering, and Applications* (Taylor & Francis, London), 71–93.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.

Borrell (2016) Forecast says SEO-related spending will be worth $80 billion by 2020. Accessed August 4, 2018, http://searchengineland.com/forecast-says-seo-related-spending-will-worth-80-billion-2020-247712.

Broder A (2002) A taxonomy of web search. *ACM SIGIR Forum* 36(2): 3–10.

Büschken J, Allenby GM (2017) Sentence-based text analysis for customer reviews. *Marketing Sci.* 35(6):953–975.

Chang J, Blei DM (2009) Relational topic models for document networks. van Dyk D, Welling M, eds. *Proc. 12th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, Vol. 5 (PMLR), 81–88.

Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Proc. 23rd Annual Conf. Neural Information Processing Systems* (Curran Associates, Red Hook, NY), 288–296.

Diebolt J, Ip EH (1995) A stochastic EM algorithm for approximating the maximum likelihood estimate. Technical report, Sandia National Laboratories, Livermore, CA.

Dzyabura D (2013) The role of changing utility in product search. Working paper, New York University, New York.

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Statist. Comput.* 24(6): 997–1016.

Gelman A, Meng XL (1996) Model checking and model improvement. Gilks WR, Richardson S, Spiegelhalter D, eds. *Markov Chain Monte Carlo in Practice* (Springer, New York), 189–201.

Geweke J (1991) *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*, Vol. 196 (Federal Reserve Bank of Minneapolis, Research Department, Minneapolis, MN).

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Ghose A, Ipeirotis P, Li B (2013) Surviving social media overload: Predicting consumer footprints on product search engines. Working paper, New York University, New York.

Green PE, Srinivasan V (1978) Conjoint analysis in consumer research: Issues and outlook. *J. Consumer Res.* 5(2):103–123.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(suppl 1):5228–5235.

Hornik K, Grün B (2011) Topicmodels: An R package for fitting topic models. *J. Statist. Software* 40(13):1–30.

Hotchkiss G, Alston S, Edwards G (2005) *Google Eye Tracking Report: How Searchers See and Click on Google Search Results* (Enquiro Search Solutions). Accessed August 4, 2018, https://searchengineland.com/figz/wp-content/seloads/2007/09/hotchkiss-eye-tracking-2005.pdf.

Jansen BJ, Booth D, Smith B (2009) Using the taxonomy of cognitive learning to model online searching. *Inform. Processing Management* 45(6):643–663.

Jansen BJ, Booth DL, Spink A (2007) Determining the user intent of web search engine queries. *Proc. 16th Internat. Conf. World Wide Web* (ACM, New York), 1149–1150.

Jansen BJ, Booth DL, Spink A (2008) Determining the informational, navigational, and transactional intent of web queries. *Inform. Processing Management* 44(3):1251–1266.

Jansen BJ, Spink A, Pfaff A (2000) Web query structure: Implications for IR system design. *Proc. 4th World Multiconference on Systemics, Cybernetics and Informatics* (IIIS, Orlando, FL), 169–176.

Jeziorski P, Segal I (2010) What makes them click: Empirical analysis of consumer demand for search advertising. Working paper, Johns Hopkins University, Baltimore, MD.

Jurafsky D, Martin JH (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 1st ed. (Prentice Hall, Upper Saddle River, NJ).

Jurafsky D, Martin JH (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. (Prentice-Hall, Upper Saddle River, NJ).

Kamvar M, Baluja S (2006) A large scale study of wireless search behavior: Google mobile search. Grinter R, Rodden T, Aoki P, Cutrell E, Jeffries R, Olson G, eds. *Proc. SIGCHI Conf. Human Factors Comput. Systems* (ACM, New York), 701–709.

Kihlstrom RE, Riordan MH (1984) Advertising as a signal. *J. Political Econom.* 92(3) 427–450.

Kim JB, Albuquerque P, Bronnenberg BJ (2010) Online demand under limited consumer search. *Marketing Sci.* 29(6):1001–1023.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Liu J, Toubia O (2018) Search query formation by strategic consumers. Working paper, Hong Kong University of Science and Technology, Hong Kong.

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval*, Vol. 1 (Cambridge University Press, Cambridge, UK).

Mimno D, McCallum A (2008) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. McAllester D,

Myllymaki P, eds. *UAI'08 Proc. 24th Conf. Uncertainty Artificial Intelligence* (AUAI Press, Arlington, VA) 411–418.

Narayanan S, Kalyanam K (2015) Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Sci.* 34(3):388–407.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Nielsen SF (2000) The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* 6(3):457–489.

Pirolli PL (2007) *Information Foraging Theory: Adaptive Interaction with Information* (Oxford University Press, New York).

Sanasam RS, Murthy HA, Gonsalves TA (2008) Determining user's interest in real time. *Proc. 17th Internat. Conf. World Wide Web* (ACM, New York), 1115–1116.

Shen Y, Yan J, Yan S, Ji L, Liu N, Chen Z (2011) Sparse hidden-dynamics conditional random fields for user intent understanding. *Proc. 20th Internat. Conf. World Wide Web* (ACM, New York), 7–16.

Shi SW, Trusov M (2013) The path to click: Are you on it? Working paper, Santa Clara University, Santa Clara, CA.

Sievert C, Shirley KE (2014) LDAvis: A method for visualizing and interpreting topics. *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces* (Association for Computational Linguistics, Stroudsburg, PA), 63–70.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. *J. Royal Statist. Soc. B* 64(4): 583–639.

Statista (2017) Paid search advertising expenditure worldwide from 2015 to 2017. Accessed August 6, 2018, https://www.statista.com/statistics/267056/paid-search-advertising-expenditure-worldwide/.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.

Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Sci.* 35(3):405–426.

Varian HR (2007) Position auctions. *Internat. J. Indust. Organ.* 25(6): 1163–1178.

Wallach HM, Mimno DM, McCallum A (2009a) Rethinking LDA: Why priors matter. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Proc. 23rd Annual Conf. Neural Information Processing Systems* (Curran Associates, Red Hook, NY), 1973–1981.

Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009b) Evaluation methods for topic models. *Proc. 26th Annual Internat. Conf. Machine Learn.* (ACM, New York), 1105–1112.

Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. Apte C, Ghosh J, Smyth P, eds. *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 448–456.

Wang P, Berry MW, Yang Y (2003) Mining longitudinal web queries: Trends and patterns. *J. Amer. Soc. Inform. Sci. Tech.* 54(8):743–758.

Wu W-C, Kelly D, Sud A (2014) Using information scent and need for cognition to understand online search behavior. *Proc. 37th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 557–566.

Yang L, Toubia O, De Jong MG (2015) A bounded rationality model of information search and choice in preference measurement. *J. Marketing Res.* 52(2):166–183.

Zhang Y, Moe WW, Schweidel DA (2017) Modeling the role of message content and influencers in social media rebroadcasting. *Internat. J. Res. Marketing* 34(1):100–119.