Marketing Science Institute Working Paper Series 2013
Report No. 13-106

Consumer Click Behavior at a Search Engine: The Role of Keyword Popularity

Kinshuk Jerath, Liye Ma, and Young-Hoon Park

Report Summary

When users search keywords at Web search engines, they are presented with both organic and sponsored links pointing to websites relevant to their search queries. Their subsequent click behavior is of great interest to advertising firms, search engines, and other practitioners and researchers who want to understand how users search for information on the Web.

In this report, Kinshuk Jerath, Liye Ma, and Young-Hoon Park address the following questions: How do consumers click on the links in the organic and sponsored lists presented after a keyword search? How does click behavior on the two lists vary across keywords? Are there systematic patterns in variations across keywords? Which characteristics of keywords can help to inform these patterns? What is the pattern of heterogeneity across searches by consumers?

Using a dataset obtained from a search engine, they analyze over 1.5 million user searches for multiple keywords and uncover robust patterns in consumer click behavior on a search results page. Using this data, they are able to paint a more complete picture of user activity on the search results page as compared to previous studies which typically use data from one advertising firm.

The authors find that, at the aggregate level, consumers' click activity after a keyword search is quite low (the modal number of clicks being one), and is concentrated on the organic list (with nearly 95% of clicks being on organic links). Interestingly, however, they find that there is significant variation in these metrics across keywords.

Their analysis reveals that consumers can be classified into different segments which can be interpreted as corresponding to different stages of consumer involvement with the topic they are searching about or the product they want to purchase. Specifically, there are low-involvement consumers and high-involvement consumers, with the latter generating more clicks per search and a larger fraction of sponsored clicks than the former. Furthermore, segments representing low-involvement consumers are composed of those who largely search more-popular keywords, and vice versa.

This study helps in developing insights into users' click behavior after a keyword search, which can be useful for advertisers. For instance, one implication of these results is that keyword popularity is an important determinant of consumers' click behavior—consumers searching more-popular keywords focus relatively more on the organic results, while consumers searching less-popular keywords focus relatively more on sponsored results. The latter are, therefore, more targetable by sponsored search advertising. This indicates that firms may want to focus their sponsored search advertising efforts on less-popular keywords, and focus their search engine optimization efforts on more-popular keywords.

Kinshuk Jerath is Assistant Professor of Marketing at the Tepper School of Business, Carnegie Mellon University. Liye Ma is Assistant Professor of Marketing at the Robert H. Smith School of Business, University of Maryland. Young-Hoon Park is AMOREPACIFIC Professor of Management and Associate Professor of Marketing at the Samuel Curtis Johnson Graduate School of Management, Cornell University. All authors contributed equally and are listed in alphabetical order.

Acknowledgments
The authors wish to thank the company, which wishes to remain anonymous, that provided the data used in this study, and Eric Bradlow, Pete Fader, Mingyu Joo, Carl Mela, Wendy Moe, Ken Wilbur, Song Yao, and Yi Zhu for their valuable comments.

Consumers use online search engines as tools to start their search for information on different topics on the World Wide Web. Examples of popular search engines include Google, Yahoo! and Bing in many countries worldwide, Yandex in Russia, Baidu in China, and Daum and Naver in Korea. When a user searches using a keyword on a search engine, she is typically presented with two lists of search results pointing to web pages relevant to her search query: a list of "organic" results, and a list of "sponsored" results.[1] The list of organic results is generated by a search engine from its proprietary database constructed by crawling and indexing billions of web pages, and using algorithms to determine the relevance of the content on a web page to the consumer's search query. The organic links often point to different types of related content, such as general information pages on the topic (e.g., pages from Wikipedia), news, blogs, images and videos. The list of sponsored links is determined using online auctions run by the search engine, where advertisers bid to be placed in response to queries by consumers. This type of advertising, called "sponsored search" or "paid search" advertising, allows firms to deliver targeted advertisements to consumers, since consumers self-identify their interest in a certain topic by searching a related keyword. Consumers, therefore, typically find both lists of results to be closely relevant to their queries (e.g., Greenspan 2004, Jansen 2007), and click on links in one or both lists to access content to satisfy their information requirements.

When presented with lists of organic and sponsored links in response to their search queries, how do consumers respond in terms of clicks on both types of links? How does click behavior on the two lists vary across keywords? Are there systematic patterns in variations across keywords? Which characteristics of keywords can help to inform these patterns? What is the pattern of heterogeneity across searches by consumers, and how can this heterogeneity be explained? With Internet users depending heavily on search engines to find information on the Web, it is crucial for advertising firms, researchers, as well as search engines, to obtain answers to the above questions. In this paper, we take a step in this direction.

The commercial success of sponsored search advertising in the last decade (eMarketer 2011)

has motivated a large body of academic work studying its various aspects. This includes both theoretical work (e.g., Edelman et al. 2007, Varian 2007, Katona and Sarvary 2010, Desai et al. 2011, Jerath et al. 2011) and empirical work (e.g., Ghose and Yang 2009, Chan and Park 2009, Yang and Ghose 2010, Yao and Mela 2011, Agarwal et al. 2011, 2012, Goldfarb and Tucker 2011, Rutz and Bucklin 2011, Joo et al. 2012). This literature primarily focuses on which keywords advertisers should bid on, what their bidding strategies should be, and how advertisers can improve the performance (in terms of click-through and conversion rates) of their sponsored advertisements. Some empirical papers have studied joint clicking behavior on sponsored and organic links. Specifically, Yang and Ghose (2010) and Agarwal et al. (2012) empirically study how the presence in the organic listing of an advertising firm's own links and of competitors' links influences click and conversion behavior for the focal firm's sponsored ads, and vice versa. Broadly speaking, both studies find complementarities between click-through rates on firms' organic and sponsored links.

While existing studies inform us on how the organic listing may influence click-through behavior for a firm's sponsored ads, they are conducted from the perspective of one single advertiser. These studies use data sourced from a single advertising firm and, therefore, lack data on clicks on sponsored and organic links of other entities on the search results page. In other words, they do not have sufficient data to give a comprehensive picture of user activity on the search results page. In this study, we use data obtained from a search engine, and have information on clicks on the *full* lists of sponsored and organic links presented after a keyword search. Using this data, we are able to paint a more complete picture of user activity on the search results page.[2]

We analyze data on approximately 1.63 million keyword searches over a one-month period for 120 keywords. For each search, we observe the numbers of clicks on the organic and sponsored lists. We model the click counts on both the organic and the sponsored lists, incorporating both observed and unobserved heterogeneity at the keyword level and also at the search instance level. From our analysis, we obtain a number of interesting and important

insights into consumers' click behavior at the search engine.

We find that, at the aggregate level, consumers' click activity after a keyword search is quite low — calculated across all searches, the average number of clicks is approximately 1.19, and the modal number of clicks is one. This result indicates that the amount of online search conducted by consumers through the results page of a keyword search, as measured by the number of websites visited, is very limited. Interestingly, this result resonates with the results of a previous study by Johnson et al. (2004), who report that the amount of online search conducted by consumers across websites is also very limited. While the scope of the two studies is somewhat different, they lead to the same general implication regarding online search behavior by consumers — a large majority of consumers do very little online search.[3] We also find that, at the aggregate level, consumers' click activity after a keyword search is concentrated on the organic results, with nearly 95% of total clicks across all searches being on organic links.

Interestingly, however, we find that there is substantial variation in these metrics across keywords. The average number of clicks per search varies across keywords between a minimum of 0.47 and a maximum of 3.67, while the average share of organic clicks varies across keywords between a minimum of 80.58% and a maximum of 99.73%. An interesting question here is: Which keyword characteristics can serve as good indicators of consumer response after a keyword search?

Previous studies have typically studied keyword characteristics such as whether the search phrase includes the name of a brand or a retailer, the length of the search phrase, etc. (e.g., Ghose and Yang 2009, Yang and Ghose 2010, Agarwal et al. 2011, Rutz and Bucklin 2011, Rutz and Trusov 2011, Rutz et al. 2011, 2012). Other studies (Rutz et al. 2012) use the semantic characteristics of the search phrase, where the semantic characteristics are determined using managerial knowledge of the business domain. These keyword characteristics are inherent to the keyword searched, i.e., they can be directly determined from the keyword. The studies mentioned above find these characteristics to be correlated with click and con-

version behavior. In our study, we incorporate such keyword characteristics (specifically, we include whether the search phrase includes the name of a brand or a retailer, and the length of the keyword; we do not include semantic characteristics) to maintain consistency with previous work. In addition, we include a new type of characteristic of a keyword — its degree of popularity. We determine the popularity score of a keyword based on its search volume, with the most-searched keyword having a "popularity rank" of 1. We call popularity rank a "new" type of characteristic because, unlike characteristics considered in previous papers, it cannot be determined by inspecting the keyword itself. In fact, it depends on how many search engine users searched the focal keyword relative to other keywords.

Interestingly, we find that the popularity score of a keyword plays a significant role in determining the click behavior of consumers. Specifically, we find that, for less popular keywords both the number of clicks per search and the share of sponsored clicks are larger as compared to more popular keywords. Furthermore, we find that different consumers have different click behavior after a keyword search, and their click patterns can be correlated with keyword popularity.

To understand this, note that organic links primarily lead a user to information-based web pages while sponsored links primarily lead a user to commercial web pages. Therefore, consumers with different click behavior on organic and sponsored links are trying to obtain different types of information, i.e., they arrive at the search engine with different intents, even if they search the same keyword. Some consumers use the search engine largely with the intent of obtaining general information about the topic of the keyword and therefore primarily click organic links; these are low-involvement consumers who may be conducting a casual search to obtain some general information. Other consumers use the search engine with the intent of gaining more detailed information on a product, possibly to carefully evaluate and subsequently purchase a product, and therefore generate more clicks and focus more on sponsored links; these are higher-involvement consumers who may be closer to making a purchase.

Our analysis reveals that consumers can be classified into different segments which can be interpreted as corresponding to different stages of consumer involvement with the topic they are searching about or the product they want to purchase. Specifically, we find that more popular keywords are searched more by lower-involvement consumers, and vice versa. Interestingly, this result resonates with the results in Moe (2003), which shows that different consumers conduct online activity in different stages of the purchase process — some consumers are simply browsing the Web, others are searching for specific information about products, while others are very close to making a purchase. We also find that the overall number of the lower-involvement consumers is significantly larger than the overall number of higher-involvement consumers. In summary, a main take-away from our paper is that the new dimension that we identify to characterize keywords, namely their popularity, is a key determinant of the number of clicks as well as the share of sponsored clicks after a keyword search.

The rest of this paper is organized in the following manner. In the next section, we provide an overview of our data and conduct exploratory data analysis to build initial insights into the interplay between click behavior on organic and sponsored links. Following this, we develop and estimate our formal model, and discuss the results and insights we obtain. We then check the robustness of our results in two different ways. We conclude with a discussion of the implications of our research and directions for future work.

DATA OVERVIEW AND EXPLORATORY ANALYSIS

In this section, we describe the data used for this research and present summary statistics on consumers' click-through behavior on sponsored and organic listings on the search results page. The patterns that we identify here assist in structuring our formal model.

We obtained a dataset of search advertising from a leading search engine firm in Korea. When a consumer searches a keyword at the search engine, she is presented with a list of sponsored links paid for by advertisers and a list of organic links chosen by the search engine. We observe which sponsored ads are displayed in response to the consumer's search query,

and which sponsored ads and organic links the user clicked. We do not have information on the full list of organic links displayed to the consumer; however, we have data on how many organic links the user clicked. We thus have data at the individual level on the number of clicks that she made on the sponsored and organic lists. In contrast, previous papers that study joint consumer behavior on sponsored and organic listings (e.g., Yang and Ghose 2010, Agarwal et al. 2012) typically have data only on whether a firm's own links were clicked or not, i.e., they lack data on the consumer's activity for the full list presented to her, because their data source is a single advertiser rather than the search engine. Note, however, that we do not have data on post-click conversion behavior.

The search engine we obtained our data from uses the following page layout when returning search results for a consumer search query. A list of sponsored ads is placed at the top of the results page, with a maximum of five ads displayed. The search engine decides which sponsored links to display, and their ordering, based on a second-price position auction. A list of organic links is placed below the list of sponsored links. The organic links are typically grouped based on the source of the content (e.g., news, blogs, images and videos), and ordered using a proprietary metric based on the relevance of the content to the keyword and the popularity of the link being displayed. Our data provider noted that there is negligible overlap between the links displayed in the organic and sponsored listings because the organic links displayed are chosen from a proprietary database consisting of data from blogs, cafes (i.e., online communities run by the portal associated with the search engine) and a knowledge database where online users post questions and other users provide answers. While collecting this dataset, websites of commercial manufacturers and sellers (i.e., the primary advertisers for sponsored results) are explicitly excluded from organic search results. On the search results page, sponsored and organic results are clearly demarcated from each other. Note that the layout used by the search engine is similar to the layout used by the major search engines in the U.S. market (such as Google, Yahoo!, and Bing), which typically display up to three sponsored links on the top of the results page followed by organic links, and

the remaining sponsored links (if any) on the right-hand side of the page.[4]

The search engine provided us data on search activity for 1,200 keywords over the one-month period (28 days) of February 2011. The keywords considered in this research were chosen and provided to us by the search engine. These keywords represent products and services for which the search engine expects consumers to be relatively active, and therefore firms also advertise on these keywords. Given keywords that pass this criterion, the search engine provided keywords to ensure significant variation in keyword search volume. Note that a "keyword" used in a query may be a single word or a phrase of a few words. The total number of search instances for 1,200 keywords add up to over 30 million. This is a prohibitively large dataset given the complexity of estimation of the model we use. Therefore, for our research, we sample 120 keywords from the 1,200 keywords uniformly at random.[5]

For each search instance in the data, the search engine records the IP address that the search instance originated from. In case of multiple search instances originating from the same IP address, the search engine has no way of knowing whether these searches were done by the same individual or different individuals. Given this limited information scenario, at one extreme, we can assume that all search instances are from different individuals, even if the IP is the same. At the other extreme, we can assume that all search instances associated with the same IP are from the same individual user. However, the market research of the search engine shows that a large fraction of IPs represents Internet cafes, offices and other such establishments where different individuals will be associated with one IP. In our opinion, neither one of the extreme assumptions seems appropriate. To resolve this issue, we randomly sample exactly *one* search instance per IP. This solution ensures, with high certainty, that no more than one search instance per individual is in the estimation data. This avoids making either of the two extreme assumptions above.[6]

We list the 120 keywords, along with their search volumes (after processing the data as described above), in Appendix A. This new dataset contains 1,631,336 million total searches across the 120 keywords. On average, there are 13,595 search queries per keyword, ranging

from a minimum of 1,241 to a maximum of 278,458 search queries. An average of 4.39 ads are displayed per search, with a relatively small standard deviation of 1.21.

From exploratory analysis of the data, we observe that consumer activity after a keyword is searched is quite limited. On average, users click only 1.19 links on the search results page. Table 1 shows the distribution of the number of links clicked. Most consumers click zero, one or two links, although a small number of consumers have much higher click counts of seven or more. Note that the modal number of clicks is one. The large fraction of users with zero clicks is consistent with previously-reported figures (e.g., Jansen and Spink 2009).

Tables follow References

We next explore patterns in joint click behavior on the organic and sponsored listings. For a given keyword, some searches do not lead to any clicks at all, others lead to clicks on either organic or sponsored listings, and yet others lead to clicks on both listings. We find that the sponsored listing, in aggregate, accounts for only 5.06% of the clicks, which is a small number. However, there is large variation in the share of sponsored clicks across keywords. To assess this variation, we calculate for each keyword the percentage of sponsored clicks it obtains (across all the times it was searched in the 28 days for which we have data). The maximum, median and minimum percentages are 19.42%, 8.91% and 0.27%, respectively. We then rank all keywords in descending order of percentage of sponsored clicks obtained and find that in the top, second, third and bottom quartiles, sponsored clicks account for 14.72%, 10.58%, 7.66% and 3.37% of total clicks, respectively. Interestingly, these quartiles account for 6.44%, 9.48%, 16.55% and 67.53% of the data by search volume, respectively; this skewness in search volume explains why the overall percentage of sponsored clicks in the data is small.[7]

To assess possible associations in consumers' click propensities, we calculate the correlation between click-through rates at both listings (calculated as the number of clicks on that listing divided by the search query volume for the keyword). The simple correlation across

the 120 keywords is 0.74 , which seems to suggest considerable dependence by consumers in clicking on both organic and sponsored lists.

Using individual-level data, we next analyze dual-click behavior in Table 2. Because of the low intensity of consumer activity, it is informative to present summary statistics conditional on observing at least one click by a consumer. (Note that, conditional on at least one click, users click an average of 1.67 links per search.) We categorize consumers based on the number of links they clicked on in each list. From Table 2, we see that consumers primarily concentrated their search activity on the organic list. Approximately 65.8% of the consumers clicked on exactly one organic link, and approximately 4.4% of the consumers clicked on exactly one sponsored link. Approximately 4.5% of consumers engaged in dual-click behavior, i.e., they clicked on at least one link from each list; this suggests that, in aggregate terms, search activity using both sponsored and organic listings is quite limited. Interestingly, this insight is different from what the correlation between the click-through rates at both listings reported above might suggest at a first look.

To summarize, our exploratory analysis indicates that, on average, consumer response on the search results page after a keyword is searched is quite limited both in terms of click-through rates and dual-click behavior. However, there is also heterogeneity in click activity. Informed by these empirical findings, we proceed towards building our statistical model.

MODEL DEVELOPMENT

Our objective is to model the number of clicks by a user on the organic and sponsored lists that she is presented with after a keyword search. Each search instance is associated with a user. Different users may arrive at the search engine with different purposes in mind, even if they are searching the same keyword. Some users search a keyword with the intent of gathering some initial information regarding a topic (represented by the keyword they search), while others are already knowledgeable about the topic and are looking for specific information. Consumers who are actively looking to satisfy a specific information need and are therefore more committed to searching can be expected to click more than

those doing a general search. Other users may be inclined to purchase a certain product and may be looking for online or offline sellers of the product, and these users can be expected to click more on the sponsored list as compared to the organic list. In general, users in different search instances are expected to have different click behavior, not only in terms of the total number of clicks but also in terms of the tendency to click sponsored versus organic links. Therefore, we allow for search instance-level heterogeneity in the model. We note that our data processing (sampling one search per IP) practically ensures that we do not have consumers in the data who conduct multiple searches. Therefore, while we can account for unobserved search instance-level heterogeneity in our model, we refrain from making inferences about consumers at the individual level. For efficiency, from now on, we call "consumer in search instance i" as "user i" or "consumer i."

We posit that consumers who search the keywords come from several different segments (Kamakura and Russell 1989). Consumers in different segments will have different behaviors on both the overall click propensity and the propensity to click sponsored or organic links. Furthermore, different keywords may draw users from the different segments in different proportions. We assume that the segment s of user i searching keyword k is a random draw from a multinomial distribution with probabilities given by the vector $\pi_k = (\pi_{k,1}, .., \pi_{k,S})$, where $\pi_{k,s}$ is the probability of being in segment s, and $\Sigma_{s=1}^{S} \pi_{k,s} = 1$. Note that this probability vector is specific to each keyword, as we account for the possibility that different keywords attract consumers from different segments in different proportions.

After searching keyword k, user i clicks on organic and sponsored links. Let $y_{ki} \in \mathbb{Z}_{\geq 0}$ denote the total number of links clicked by consumer i at keyword k (including clicks on both sponsored and organic links). We assume that the number of clicks after keywork k is searched follows a Poisson distribution with rate parameter $\lambda_{ki,s}$:

$$y_{ki} \sim \text{Poisson}(\lambda_{ki,s}),$$

where $\lambda_{ki,s}$ is a click-propensity parameter that captures consumer i's mean click propensity

for keyword k given that she belongs to segment s. We model $\lambda_{ki,s}$ as follows:

$$\ln(\lambda_{ki,s}) = \beta^{\lambda}_{k,s} + \beta^{\lambda}_{\text{POP}} \cdot \ln(\text{Popular}_k) + \beta^{\lambda}_{X} X_k + \beta^{\lambda}_{Z} Z_i. \quad (1)$$

With respect to the covariates in Equation (1), Popular_k is a measure of the popularity of keyword k and is defined as the rank of keyword k on the basis of the search query volume during the data period, with the most-searched keyword ranked at the top as 1. Therefore, a larger value of Popular_k indicates that the keyword is *less* popular. X_k is a vector of keyword-specific covariates. Through X_k, we include three important observed keyword-specific characteristics, which capture the nature of the keyword. These include: (1) whether the keyword has retailer-specific information (i.e., whether a seller/retailer name appears in the query), (2) whether the keyword has brand-specific information (i.e., whether a brand name appears in the query), and (3) the length (number of words) of the keyword. Based on these factors, we construct three keyword-specific characteristics denoted by Retailer_k, Brand_k, and Length_k. The first two variables are coded as dummy variables, while the third is a natural number. We employ these keyword-specific covariates to control for observed heterogeneity across keywords, which is consistent with prior work in this area. Z_i includes covariates specific to the instance of the search by consumer i. We incorporate two covariates here. First, we include Num_Sponsored_i, which is a whole number, and denotes the number of sponsored links displayed after the keyword search is conducted (the number of available sponsored links can be expected to influence the number of sponsored links clicked, and therefore the overall number of clicks as well). Second, we include Weekend_i which is a dummy variable, and denotes a weekend indicator to account for day-of-the-week effect.[8]

With respect to the parameters in Equation (1), the parameter $\beta^{\lambda}_{k,s}$ represents the baseline click propensity for a consumer in latent segment s after searching keyword k. Consumers in different segments have different keyword-segment specific intercepts, $\beta^{\lambda}_{k,s}$, and this captures heterogeneity across consumers in their tendency to click after searching keyword k. The parameter $\beta^{\lambda}_{\text{POP}}$ indicates how click propensity changes as a function of the keyword popular-

ity: If $\beta^{\lambda}_{\text{POP}}$ is negative, the number of clicks (per search) is greater for more popular (higher search volume) keywords, while if $\beta^{\lambda}_{\text{POP}}$ is positive, then the number clicks (per search) is lesser for more popular keywords. β^{λ}_{X} is a vector of three coefficients measuring the impact of Retailer_k, Brand_k, and Length_k, respectively, on the propensity to click. β^{λ}_{Z} is a vector of two coefficients measuring the impact of Num_Sponsored_i and Weekend_i, respectively, on the propensity to click.

We now incorporate dual-click behavior in the model (i.e., whether the user clicks on a sponsored link or an organic link). Let y^{O}_{ki} and $y^{S}_{ki} = y_{ki} - y^{O}_{ki}$ denote the number of organic links and sponsored links clicked by consumer i after searching keyword k, respectively. We assume that for each click after a keyword search, there is probability $p_{ki,s}$ that the click will be on a sponsored link, i.e., we assume a Bernoulli process. Note that our assumptions of a Poisson process followed by a Bernoulli process imply that y^{O}_{ki} and y^{S}_{ki} also follow Poisson distributions with rate parameters $\lambda_{ki,s} \cdot (1 - p_{ki,s})$ and $\lambda_{ki,s} \cdot p_{ki,s}$, respectively.

Similar to Equation (1), we parametertize $p_{ki,s}$ as:

$$\text{logit}(p_{ki,s}) = \beta^{p}_{k,s} + \beta^{p}_{\text{POP}} \cdot \ln(\text{Popular}_k) + \beta^{p}_{X} X_k + \beta^{p}_{Z} Z_i. \tag{2}$$

In Equation (2), the parameter $\beta^{p}_{k,s}$ represents the baseline probability for clicking a sponsored link after searching keyword k for a consumer in latent segment s. This keyword-segment specific intercept captures heterogeneity across consumers in their tendency to search for information in the sponsored versus organic lists after searching keyword k. The parameter β^{p}_{POP} indicates how click probability for sponsored links changes as a function of the keyword popularity, as measured by Popular_k. If β^{p}_{POP} is positive, then the likelihood to click a sponsored link increases with keyword popularity, and vice versa. β^{p}_{X} is a vector of three coefficients measuring the impact of the components of X_k on the probability to click a sponsored link, and β^{p}_{Z} is a vector of two coefficients measuring the impact of the components of Z_i on the probability to click a sponsored link.

We adopt a hierarchical Bayesian framework and assume that the keyword-segment spe-

cific intercepts for click and sponsored propensity for each keyword-segment pair are random draws from segment-specific normal distributions given by:

$$\beta_{k,s}^{\lambda} \sim N(\bar{\beta}_s^{\lambda}, (\sigma_s^{\lambda})^2) \text{ and } \beta_{k,s}^{p} \sim N(\bar{\beta}_s^{p}, (\sigma_s^{p})^2), \tag{3}$$

where $\bar{\beta}_s^{\lambda}$ and $\bar{\beta}_s^{p}$ are the respective population-level means, and $(\sigma_s^{\lambda})^2$ and $(\sigma_s^{p})^2$ are the respective population-level variances. We note that, except for the baseline parameters (i.e., $\beta_{k,s}^{\lambda}$ and $\beta_{k,s}^{p}$) which are segment specific, we assume all parameters to be population specific to maintain simplicity of interpretation of the results.

In summary, click behavior on the search results page is governed by two components of the model. The first component determines the overall propensity to click after a keyword is searched, and the second component determines the likelihood to search for information in the sponsored versus organic listings. Furthermore, our model accounts for: (1) observed heterogeneity in keywords (via keyword popularity and other keyword characteristics), (2) observed heterogeneity in consumers (via characteristics of the search instances), and (3) unobserved heterogeneity among consumers (via latent segments). By allowing different keywords to attract consumers from different segments, we can infer patterns in the composition of consumers searching different keywords.

ESTIMATION AND RESULTS

Estimation

We have a total of 1,631,336 search instances for the 120 keywords we consider. To allow for a shorter estimation time, we randomly sample 20% of the consumer searches from the above; this 20% subsample contains 326,080 search instances. We compared a few summary statistics between the full data of the 120 chosen keywords and the 20% sample of these data. The means and standard deviations of total clicks, sponsored clicks, organic clicks, and other variables are all very close between the full data and the 20% sample (see Table 3). This is not surprising, considering the large number of observations in the dataset.

We adopt a Bayesian approach and use the Markov chain Monte Carlo (MCMC) method to estimate our proposed model. The details of the MCMC procedure are given in Appendix B. We draw samples from the posterior distribution of 40,000 iterations from two independent MCMC chains following a burn-in of 40,000 iterations. Our proposed model accounts for observed heterogeneity in keywords (via observed keyword characteristics), observed heterogeneity in search instances (via observed search instance characteristics), and unobserved heterogeneity among consumers (via latent segments). Furthermore, by allowing different keywords to attract consumers from different segments, our model allows the inferences in the composition of consumers searching different keywords.

Model Fit

We estimate the model proposed above with different numbers of consumer segments, ranging from two to six. To aid in model selection to determine the optimal number of segments, we use both log marginal density (LMD) and mean absolute error (MAE) in predicted total number of clicks. We report these model fit measures in Table 4. The table shows that as the number of consumer segments increases from two to four, LMD increases significantly; after four segments, however, LMD practically levels off. The table also shows that MAE is virtually identical for different numbers of segments. The results suggest that the proposed model with four segments performs well in terms of both model fit measures, and we therefore focus on this model hereafter. (The results with five or more segments are qualitatively the same, though they are somewhat cumbersome to interpret.)

As a measure of the accuracy of the model with four segments, we calculate, for each keyword, the expected number of organic and sponsored clicks over the data period and compare them with the actual number of clicks. Across the 120 keywords, mean absolute percentage errors (MAPE), weighted by search volume, are 2.33%, 1.08% and 2.09% for organic, sponsored and total clicks, respectively. MAPEs which are not adjusted by search volume are 5.83%, 2.72% and 5.07% for organic, sponsored and total clicks, respectively. These statistics provide strong evidence that the proposed model with four segments performs

well in capturing click behavior for both organic and sponsored links at the keyword level, and inspire confidence in the validity of the model.

Results

We organize the reporting of our model results, and the associated insights about user click behavior after a keyword search, into three main parts. First, we describe inferences regarding the characteristics of the four latent segments based on the inferences regarding the click behavior of consumers in the segments and the keyword loadings on these segments. Second, we describe inferences regarding the impact of keyword popularity. Third, we describe inferences regarding the impact of covariates capturing observed heterogeneity through X_k and Z_i. All parameter estimates are reported in Table 5.

Characteristics of segments. As discussed earlier, we obtain four latent segments. We first look at the keyword-segment specific intercepts for click propensity. The population-level mean estimates, denoted by $\bar{\beta}^{\lambda}_{s}, s \in \{1, 2, 3, 4\}$, are -0.128, 0.478, 1.227, and 2.124, respectively; these correspond to 0.88, 1.61, 3.41 and 8.36 average clicks per search for Segments 1, 2, 3 and 4, respectively. Keeping the same order of segments, the population-level mean estimates for the keyword-segment specific intercepts for the propensity to click sponsored links, denoted by $\bar{\beta}^{p}_{s}, s \in \{1, 2, 3, 4\}$, are -3.574, -3.362, -1.660, and -2.346, respectively; these correspond to sponsored click probabilities of 2.73%, 3.35%, 15.98%, and 8.74% for Segments 1, 2, 3 and 4, respectively. These estimates essentially imply that, in the ordering we impose, for higher-numbered segments (i.e., Segments 3 and 4, as compared to Segments 1 and 2), consumers are inclined to click more links per search, and also to click sponsored links with higher probability.

The segment descriptions above provide an interesting insight about the search behavior of consumers at a search engine. It is widely accepted in marketing that consumers move towards a purchase through a hierarchical sequence of events, from cognition (thinking, e.g., awareness, consideration) to affect (feeling, e.g., liking, preference), and ultimately conation (doing, e.g., purchase intent, purchase). These concepts were integrated into general models

of consumer behavior (Howard and Sheth 1969). For instance, the Awareness-Interest-Desire-Action model (or AIDA model) is one model that captures this multi-stage decision process phenomenon. Such models are also known as "purchase funnel" models because only a fraction of consumers proceed from one stage to the next, i.e., the consumer base narrows sequentially through the stages. Interestingly, our estimates for keyword-segment specific intercepts for click propensity and propensity to click sponsored links indicate that the characteristics of the four latent segments that we uncover are in agreement with the purchase funnel theory. Recall that we are considering commercial keywords here, and the search engine guarantees negligible overlap between the organic and sponsored lists. This indicates that consumers in higher-numbered segments, as compared to consumers in lower-numbered segments, are in more advanced stages of involvement with the product/category being searched (in terms of acquiring information on it, and potentially purchasing it) because they go through more of the results returned after the keyword search and also devote more attention to sponsored results with links to commercial websites.

For further agreement with the purchase funnel theory, we would expect to see that higher-numbered segments (which have consumers with more clicks per search and proportionally more sponsored clicks) are smaller in size since a larger number of searches by consumers are expected to be for gathering information at the general level than for detailed product/seller evaluation. Indeed, we find that, based on search volume, Segments 1, 2, 3 and 4 have relative sizes of 49.11%, 44.96%, 4.20% and 1.73%, respectively (i.e., 49.11% of all searches fall in Segment 1, 44.96% of all searches fall in Segment 2, ...).

Furthermore, our estimates show that less popular keywords in general have a larger portion of their searches by consumers from the higher-numbered segments. This can be seen from Table 6, in which we report the average percentage of consumers from each segment for the top 30 most-popular keywords, the next 30 most-popular keywords, and so on in our 120-keyword dataset. We find that as the keyword ranks increase (i.e., as the keywords become less popular), the proportions of Segment 1 and Segment 2 decrease, while the proportions of

Segment 3 and Segment 4 increase. We see a clear pattern that Segments 1 and 2, which we call the low-involvement segments, have a larger proportion of keywords with higher search volume than Segments 3 and 4, which are high-involvement segments.

Taken together, the above inferences based on our segment analysis strongly suggest that the four latent segments we uncover correspond to a range of consumers conducting searches in different stages of the purchase process — Segments 1 and 2 represent lower-involvement searchers and Segments 3 and 4 represent high-involvement searchers. Furthermore, we find that more popular keywords are searched more by consumers in early, low-involvement stages of the purchase process while less popular keywords are searched more by consumers in advanced, high-involvement stages of the purchase process. As a consequence, more popular keywords have fewer clicks per search and a larger proportion of clicks on the organic listing since most consumers searching these keywords are gathering information at a general level; in contrast, less popular keywords have more clicks per search and a larger proportion of clicks on the sponsored listing since most consumers searching these keywords are more involved and will click on more links to get more thorough information, and since they are closer to purchase, they will click on more sponsored links which are of commercial nature compared with organic links.

Effect of popularity. From the posterior means of $\beta^{\lambda}_{\text{POP}}$ and β^{p}_{POP}, we find that more popular keywords receive fewer clicks per search and receive a smaller fraction of clicks on sponsored links, which is in agreement with the analysis discussed above. An important point here is that, if we do not allow for consumer heterogeneity through segments (i.e., we only allow one segment of consumers), both these coefficients have values that are considerably larger and have the same signs, i.e., both effects are in the same direction but are stronger.[9] This leads to a very interesting conclusion — while keyword popularity can explain click behavior, a large part of the impact of popularity is through consumer selection into different segments. In other words, the fact that inclusion of multiple segments weakens the direct impact of popularity indicates that the effect of popularity on click behavior is through the

different stages of involvement of consumers.

Effect of covariates in X_k and Z_i. Looking at the other parameters for observed heterogeneity, we find that the covariates Retailer_k, Brand_k and Length_k largely have no impact on either overall click propensity after a search or the propensity to click a sponsored link. The only exception is that the click propensity is lower if the searched keyword contains retailer-specific information. Turning to the covariates specific to the search instance, we find that the number of sponsored links displayed at the time of search is positively correlated with the propensity to click sponsored links. This may be due to an agglomeration effect, i.e., more sponsored links draw greater attention from the user. We also find that a weekend search is not different from a weekday search in terms of either the overall click propensity or the propensity to click sponsored links.

Comparison with Previous Literature

A number of existing papers study the impact of observed keyword characteristics on click behavior (e.g., Ghose and Yang 2009, Yang and Ghose 2010, Agarwal et al. 2011, 2012, Rutz and Bucklin 2011, Rutz and Trusov 2011, Rutz et al. 2011, 2012). Broadly speaking, these studies find that search phrases that include a retailer name or a brand name, and search phrases that are longer, have higher click-through rates on sponsored links. We note that these papers do not consider popularity as a covariate. Our results are not directly comparable to the results in these studies because they analyze click data obtained from one single firm, while we analyze click data obtained from the search engine for the full list of links, and for keywords that are relevant to many different firms and industries. Nevertheless, it is informative to compare our results to those in the existing literature.

Interestingly, we do not find strong effects of the above observed keyword characteristics (namely, the presence of a retailer name or a brand name in the search phrase, and the length of the search phrase) on either the overall click propensity or the propensity to click sponsored versus organic links. However, we do find a strong effect of keyword popularity (i.e., the relative search volume) on both the overall click propensity and the propensity to

click sponsored versus organic links. In our data, the correlation between keyword search volumes and presence of a retailer name in the keyword is 0.182, the correlation between keyword search volume and presence of a brand name in the keyword is 0.169, and the correlation between keyword search volume and length of the keyword is 0.003; all these correlations are weak. In this light, our results indicate that keyword popularity is an important characteristic that determines click behavior after a keyword search, and this characteristic is, in general, different from the observed keyword characteristics mentioned above. (While it would be reasonable to argue that search phrases that contain the name of a retailer or a brand in them, or search phrases that are longer, are searched less, this does not imply that all keywords that are searched less have one or more of the above observed characteristics.) In this light, we note that while our results are different from results in previous papers, they are not at odds with these results; rather, they are complementary to previous results. Future research can further enhance our understanding of these issues.

ROBUSTNESS OF RESULTS

Latent Popularity Scores

In the previous section, we use observed keyword rank, i.e., keyword rank derived directly from observed keyword search volumes, as the popularity score for a keyword. The advantage of this approach is that it gives direct evidence of keyword popularity as an indicator of searchers' click tendencies. However, one could think of the popularity of a keyword as a *latent* construct, based on which the observed number of searches is determined, with some stochasticity in the outcome. In addition, observed keyword characteristics might also play a role in determining the observed number of searches of the keyword. In this section, we develop a model to estimate the latent popularity score for every keyword from the search volume data while controlling for keyword characteristics. We then use the obtained score instead of Popular_k in Equations (1) and (2) to check whether keyword popularity plays the same role as we find in the previous section.

To estimate the latent popularity score, we augment the model described earlier with the following component. Given S segments (with π_s denoting the relative size of segment $s \in \{1, \ldots, S\}$), we assume that each keyword k has a score specific to segment s, $\Theta_{k,s}$, given by:

$$\Theta_{k,s} = \theta_{k,s} + \theta_X X_k + \varepsilon_{k,s}.$$

In the above equation, $\theta_{k,s}$ represents the intrinsic, unobserved loading of keyword k on segment s, $\theta_X X_k$ controls for observed keyword characteristics and $\varepsilon_{k,s}$ is a random component. Essentially, $\theta_{k,s}$ is a measure of the latent relative popularity of keyword k among all K keywords with respect to segment s. We assume that $\varepsilon_{k,s}$ follows the extreme value distribution, which implies that, fixing the segment as s, the probability that keyword k will be searched is given by:

$$\Pr(\text{search keyword } k) = \frac{\exp(\theta_{k,s} + \alpha_X X_k)}{\sum_{l=1}^{K} \exp(\theta_{l,s} + \alpha_X X_l)}.$$

The rest of the model is exactly the same, except that in Equations (1) and (2), Popular_k is replaced with $\theta_{k,s}$.

On estimating this model, for all the parameters that are common across the two models, we obtain results that are qualitatively identical to the results presented in the previous section. This analysis strengthens our insights regarding the role of keyword popularity in determining click behavior after a keyword search. More details are available on request.

Data Pre-Processing

As explained in the data overview section, we pre-process the raw data obtained from the search engine by first randomly choosing 120 keywords from the 1,200 keywords which we were provided, and then randomly sampling exactly one search instance per IP. This essentially ensures that we do not have multiple searches by one individual in the data. Preprocessing the data in different ways gives the same insights regarding the impact of keyword popularity. We pre-process the data in two different ways and estimate our model on the resulting data. In the first alternative method of pre-processing the data, we first randomly sample one search instance per IP in the data and then randomly choose 120 keywords from

1,200 keywords. In the second alternative method of pre-processing the data, we include all search instances for all IPs and randomly sample 120 keywords from 1,200 keywords. (In this case, we treat each search instance as independent from other search instances, even if it is from the same IP address.) In both cases, we obtain results that are qualitatively identical to the results presented in the previous section. More details are available on request.

CONCLUSIONS AND DISCUSSION

In this paper, we study consumers' click behavior on the organic and sponsored results presented to them after a keyword search at an Internet search engine. We analyze data from over 1.5 million keyword searches of 120 different keywords over the span of one month. Our study is unique in several aspects. First, we analyze rich click-through data obtained from the search engine for the full list of results presented to a user after every keyword search. In contrast, previous related studies have analyzed data obtained from one single firm for clicks on its own links (i.e., data on clicks on other firms' links displayed on the search results page is absent). Second, to understand how keyword characteristics influence click behavior, previous studies have typically used observed keyword characteristics (such as the presence of a retailer name or a brand name in the search phrase, and the length of the search phrase). We use, in addition to observed characteristics, a keyword's popularity score, i.e., what is the relative popularity of a keyword in terms of search volume. We find that keyword popularity is strongly correlated with the total clicks and the proportion of sponsored clicks that are generated after a keyword search. Third, unlike previous studies, we include observed and unobserved search instance-level heterogeneity (which, in our case, is equivalent to consumer-level heterogeneity), and find that unobserved heterogeneity is especially important in understanding patterns in click activity after a keyword search.

Our results show that average click activity after a keyword search is quite low (the modal number of clicks being one), and is concentrated on the organic list (with nearly 95% of clicks being on organic links). However, there is significant heterogeneity across users conducting the different searches. We find that users can be grouped into latent segments;

this aids in understanding consumers' click behaviors because the segments can be considered as representing consumers in different stages of involvement with the topic or product they are searching about. Specifically, there are low-involvement consumers and high-involvement consumers, with the latter generating more clicks per search and a larger fraction of sponsored clicks than the former. Furthermore, segments representing low-involvement consumers are composed of those who largely search more popular keywords, and vice versa.

Our study helps in developing insights into users' click behavior after a keyword search, which can be useful for advertisers. For instance, one implication of our results is that keyword popularity is an important determinant of consumers' click behavior — consumers searching more popular keywords focus relatively more on the organic results, while consumers searching less popular keywords focus relatively more on sponsored results. The latter are, therefore, more targetable by sponsored search advertising. This indicates that firms may want to focus their sponsored search advertising efforts on less popular keywords, and focus their search engine optimization efforts on more popular keywords (for instance, choosing the website content to increase the website's relevance to specific keywords; Berman and Katona 2011). In addition to the above, the insights that we develop into consumers' click behavior can also help search engines to design better responses to consumer queries, and therefore better serve both search engine users and advertisers.

We use a large and rich dataset to identify salient patterns in how users click on lists of organic and sponsored links after a keyword search. Future work can build on our findings by replicating our analysis on similar datasets obtained from other sources, which can possibly lead to empirical generalizations regarding post-search consumer click activity. Future work can also address certain shortcomings of our research. First, the click behavior of consumers will depend on the relevance of the results presented to the keyword searched. Given our data, we are unable to address this aspect. Richer data is needed to explicitly incorporate relevance into the model; for instance, we may need data on the identities of all firms that are displayed in the organic and sponsored lists, the ad copies used, the landing

pages that the user is directed to, etc. Second, we show that keyword popularity is an important indicator of searchers' click tendencies and provide arguments that consumers in the different segments that we identify are in different stages of involvement with respect to the relevant search based on their click behavior. However, we stop short of making causality arguments. Further research can control for stages of consumer involvement more carefully by running experiments, possibly in the manner of Lambrecht et al. (2011) and Lambrecht and Tucker (2012), to infer causality. Third, relative popularity of keywords will change over time. It will be interesting to obtain temporal data to study the length of time for which keyword popularity is stable and how this varies across keywords. Finally, availability of data on post-click conversion rates can further enhance our understanding of consumer behavior after a keyword search. We hope that our study can motivate further research in the above directions.

FOOTNOTES

[1]Throughout the paper, we use "keyword," "query," "search query," "search phrase," etc. interchnageably. A "keyword" may be a phrase with more than one words in it.

[2]An early paper in this spirit, which uses data from a search engine, is Pass et al. (2006). However, we focus on different questions in this paper.

[3]Some previous studies (e.g., Goodman 2010) have reported the phenomenon of "navigational searches," i.e., consumers search a keyword at a search engine simply to obtain the exact URL of a web site they want to visit; after the search, they click only on this URL. Such navigational searches may contribute to the reasons that the modal number of clicks in our data is one.

[4]In our setting, there can be more sponsored links (up to five) at the top of the search results page. This difference in the layout of the results page can lead to some difference in consumer response to the sponsored and organic lists, with more clicks expected on the sponsored links in our setting. However, with approximately 29% and 47% of keyword searches having zero and one clicks (including clicks on both sponsored and organic links), respectively, the impact of the difference in the results page layout is expected to be relatively small, and the insights we obtain from our analysis can inform us to a large extent about user click behavior on other search engines as well.

[5]An assessment using exploratory methods indicated that the dataset with the 120 sampled keywords is representative of the full dataset with 1,200 keywords.

[6]Pre-processing the data in the manner described also removes the need to incorporate individual-level effects in the model. Since the focus of this paper is not to track or analyze individual-level behavior across search instances or a set of keywords searched, we believe that sampling in this way is an appropriate method of pre-processing the data.

[7]According to research based on 28 million people in the UK, making a total of 1.4 billion search queries during June 2011, paid search only accounts for 6% of total clicks from search engines versus organic search at 94% of clicks (GroupM 2012). Our number of 5.06% of sponsored clicks is comparable, though slightly lower. We do not know of any previous academic study that has reported this figure (across many keywords, and accounting for all advertisers that are listed).

[8]We do not have any data on individual demographics (such as age, income, sex, etc.).

[9]We estimate a model with one segment (i.e., we force $S = 1$), while keeping everything else in the model the same. In this model, $\beta^{\lambda}_{\text{POP}}$ has the value 0.053 with the 95% credible interval being $[0.019, 0.097]$, and β^{p}_{POP} has the value 0.606 with the 95% credible interval being $[0.506, 0.719]$. The values of both coefficients are considerably larger compared to the four-segment model.

REFERENCES

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

Agarwal, Ashish, Kartik Hosanagar, and Michael D. Smith (2012), "Sponsored Search: How Organic Results Impact Sponsored Search Advertising Performance," Working Paper, University of Texas at Austin.

Berman, Ron and Zsolt Katona (2011), "The Role of Search Engine Optimization in Search Marketing," Working Paper, University of California, Berkeley.

Chan, Tat Y. and Young-Hoon Park (2009), "Position Competition in Sponsored Search Advertising," Working Paper, Washington University in St Louis..

Desai, Preyas, Shin Woochoel and Richard Staelin (2011), "The Company that You Keep: When to Buy a Competitors Keyword," Working Paper, Duke University.

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

eMarketer (2011), "US Online Ad Spending Growth by Format 2010-2015," June 2011.

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

Goldfarb, Avi and Catherine Tucker (2011), "Search Engine Advertising: Channel Substitution When Pricing Ads to Context," *Management Science*, 57 (3), 458–470.

Goodman, Eli (2010), "Navigational Search: Turn Right at the Big Chicken," *Search Engine Watch*, September 13.

Greenspan, Robyn (2004), "Searching for Balance," ClickZ, April 30.

GroupM (2012), "Evaluating the UK Search Marketing Landscape: Exposing SEO CTRs by Industy and Who Clicks on PPC," August 2012.

Howard, John A. and Jagdish N. Sheth (1969), *The Theory of Buyer Behavior,* New York, Wiley.

Jansen, Bernard J. (2007), "The Comparative Effectiveness of Sponsored and Non-sponsored Links for Web e-commerce," *ACM Transactions on the Web*, 1 (1), Article 3.

Jansen, Bernard J. and Amanda Spink (2009), "Investigating Customer Click Through Behavior with Integrated Sponsored and Nonsponsored Results," *International Journal of Internet Marketing and Advertising*, 5 (1/2), 74–94.

Jerath, Kinshuk, Liye Ma, Young-Hoon Park, and Kannan Srinivasan (2011), "A 'Position Paradox' in Sponsored Search Auctions," *Marketing Science*, 30 (4), 612–627.

Johnson, Eric J., Wendy W. Moe, Peter S. Fader, Steven Bellman, and Jerry Lohse (2004), "On the Depth and Dynamics of World Wide Web Shopping Behavior," *Management Science*, 50 (3), 299–308.

Joo, Mingyu, Kenneth C. Wilbur, and Yi Zhu (2012), "Television Advertising and Online Search," Working Paper, Ohio State University.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probablistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (4), 379–390.

Katona, Zsolt and Miklos Sarvary (2010), "The Race for Sponsored Links: Bidding Patterns for Search Advertising," *Marketing Science*, 29 (2), 199–215.

Lambrecht, Anja, Katya Seim and Catherine Tucker (2011), "Stuck in the Adoption Funnel: The Effect of Interruptions in the Adoption Process on Usage," *Marketing Science*, 30 (2), 355–367.

Lambrecht, Anja and Catherine Tucker (2012), "When Does Retargeting Work? Timing Information Specificity," Working Paper, London Business School.

Moe, Wendy W. (2003), "Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology*, 13 (1&2), 29–40.

Pass, Greg, Abdur Chowdhury and Cayley Torgeson (2006), "A Picture of Search," *Proceedings of the First International Conference on Scalable Information Systems*, May 29–June 1, 2006, Hong Kong.

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

Rutz, Oliver, Randolph E. Bucklin and Garrett P. Sonnier (2012), "A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising," *Journal of Marketing Research*, 49 (3), 306–319.

Rutz, Oliver and Michael Trusov (2011), "Zooming In on Paid Search Ads — A Consumer-Level Model Calibrated on Aggregated Data," *Marketing Science*, 30 (5), 789–800.

Rutz, Oliver, Michael Trusov, and Randolph E. Bucklin (2011), "Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?," *Marketing Science*, 30 (4), 646-665.

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

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

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

No. of clicks	0	1	2	3	4	5	6	7+
Frequency (%)	28.6	46.7	15.2	4.6	2.0	1.1	0.6	1.2

Table 1: Total Number of Clicks Per Search

		No. of clicks on sponsored list			
		0	1	2	3+
No. of clicks	0	–	1.1	0.2	0.1
on organic list	1	64.3	1.2	0.2	0.1
	2	19.8	0.8	0.2	0.1
	3+	9.9	1.3	0.5	0.3

Table 2: Distribution of Clicks (%) Conditional on At Least One Click

	Full data		20% Sample	
	Mean	Std. Dev.	Mean	Std. Dev.
Total clicks	1.19	1.45	1.19	1.46
Sponsored clicks	0.06	0.34	0.06	0.33
Organic clicks	1.13	1.33	1.13	1.34
Number of sponsored links	4.39	1.21	4.39	1.21
Weekend	0.26	0.44	0.26	0.44
Total observations	1,631,336		326,080	

Table 3: Descriptive Statistics

No. of segments	LMD	MAE
2	-494724.7	1.29
3	-491879.7	1.28
4	-491099.1	1.28
5	-490650.8	1.28
6	-490555.4	1.28

Table 4: LMD and MAE for Different Numbers of Segments

Parameters Describing Latent Segments				
	Segment 1	Segment 2	Segment 3	Segment 4
$\bar{\beta}_s^{\lambda}$	-0.128 [-0.267, 0.013]	0.478 [0.338, 0.623]	1.227 [1.105, 1.341]	2.124 [1.996, 2.248]
$(\sigma_s^{\lambda})^2$	0.529 [0.393, 0.699]	0.383 [0.287, 0.513]	0.260 [0.182, 0.361]	0.414 [0.298, 0.559]
$\bar{\beta}_s^{p}$	-3.574 [-3.884,-3.297]	-3.362 [-3.755,-2.998]	-1.660 [-1.978,-1.369]	-2.346 [-2.547,-2.139]
$(\sigma_s^{p})^2$	1.714 [1.135, 2.685]	3.216 [2.316, 4.333]	2.176 [1.494, 3.033]	0.946 [0.641, 1.329]

Coefficients for Keyword-Level Covariates	
$\beta_{\text{POP}}^{\lambda}$	0.012 [0.003, 0.020]
β_{POP}^{p}	0.299 [0.095, 0.425]

	Retailer_k	Brand_k	Length_k
β_X^{λ}	-0.168 [-0.233,-0.069]	-0.044 [-0.150, 0.027]	0.035 [-0.012, 0.087]
β_X^{p}	0.178 [-0.243, 0.556]	-0.170 [-0.497, 0.037]	-0.009 [-0.147, 0.107]

Coefficients for Search-Level Covariates		
	Num_Sponsored_i	Weekend_i
β_Z^{λ}	-0.018 [-0.029,-0.007]	-0.007 [-0.016, 0.001]
β_Z^{p}	0.413 [0.360, 0.463]	-0.007 [-0.048, 0.038]

Table 5: Parameter Estimates (values in brackets are the 95% credible intervals)

Keyword rank	Segment 1	Segment 2	Segment 3	Segment 4
1-30	56.22%	37.85%	3.81%	2.11%
31-60	54.61%	33.22%	9.26%	2.91%
61-90	56.12%	28.32%	11.54%	4.01%
91-120	49.54%	29.46%	14.98%	6.02%

Table 6: Segment Proportions for Keywords (Grouped by Keyword Popularity Rank)

APPENDIX A: LIST OF KEYWORDS

Below we list the 120 keywords we use in our estimation. The keywords are listed in decreasing order of search volume, and each keyword is followed by the number of times it was searched in the time period over which our data was collected. Note that the keywords were originally in Korean, and have been translated into English.

Hyundai Card, 278458; Lotte.com, 146869; Samsung Electronics, 90586; Lotte Department Store, 84440; Lotte Mart, 63649; LG Electronics, 34171; iPhone 4, 50088; Smartphone, 40296; iPad, 35361; Newly-released movies, 38538; North Face, 28776; LG Telecom, 30127; Shilla Duty-free Store, 24724; Bean Pole, 20811; Costco, 24712; iPad 2, 22449; Car, 20844; Nike Shoe, 19362; Lego, 15102; Adidas, 16795; Louis Vuitton Bag, 13746; Netbook, 14395; Kipling, 12659; Gucci, 11282; Travel, 12963; Computer, 9651; Travel Agency, 10543; Auto Insurance, 8778; Watch, 14848; Abercrombie, 12144; Daks, 12475; Navigation, 12499; Gucci Bag, 11149; Wallet, 11525; Luxury Bag, 10960; Travel to Jeju Island, 9733; Desk, 10303; Cyber University, 7170; Sisley, 10080; Shopping Mall for Women's Clothes, 9208; Mountain-climbing Equipment, 7182; Coach Bag, 8510; Car Rental, 15424; Zara, 9182; Mountain-climbing Clothes, 7050; Cardigan, 8432; Golf Club, 6195; Online Loan, 1933; Mountain-climing Shoe, 6559; Insurance Comparison Site, 2411; Valentine's Day Gift, 7960; Ohui, 6806; Loem, 6788; Shopping Mall by Celebrity, 7458; Auction Site, 7071; Outdoor, 6018; Wedding Dress, 5272; Bluedog, 5908; Audio, 3386; Gapyung Pension, 4503; Ziozia, 3922; Michael Kors, 5080; Shopping Mall for Men, 5099; Water Purifier, 3944; Social Commerce, 5337; Humidfier, 4767; Earring, 4838; Hot Spring, 5121; Canon Camera, 5176; Used, 8088; Wedding Information Agency, 2231; Chatelaine, 3603; Shopping Mall Ranking for Men's Clothes, 4245; Travel to Bally, 2225; Adidas Running Shoe, 3061; Curtain, 3640; Ugg Boots, 2062; Airconditioner, 3189; Kimchi, 2931; Golf Shoe, 1508; Shoe Shopping Mall, 2312; Travel to Singapore, 1873; Used Luxury, 2760; Men's Clothes, 2303; Pet Sale, 2981; Nike Bag, 3034; Travel to Japanese Hot Spring, 1661; Headband, 3126; Used Monitor, 1299; Tory Burch, 2703; Free Travel to Hong Kong, 1241; Yoga Clothes, 2736; North Face Padding, 4095; Travel to Australia, 1509; Cheap Furniture Store, 2986; Daks Wallet, 2292; Leggings, 2895; Shopping Mall for Pretty Bags, 3032; Columbia, 3215; Treadmill, 2415; Island Dining Table, 2666; Pretty Curtain Store, 2616; Bed Discount Store, 2743; Running Shoe Shop, 4891; Kitchen Appliance, 2394; Crocodile, 2282; Women's Suit, 2403; LED Lighting, 2126; Summer Clothes Shopping Mall, 2199; Coffee Bean, 3828; iPod Touch 4, 1926; Bang Bang, 2383; Doosan Otto Shopping Mall, 2273; Bicycle Equipment, 1957; Notebook Bag, 2053; Handphone Case, 2144; Luxury Style Women's Clothes, 1983; Shopping Mall for Women's (30s) Clothes, 5709; Used Motorcycle, 3834; Livingroom Interior, 2074.

APPENDIX B: DETAILS OF MCMC ESTIMATION PROCEDURE

We estimate our model using a Markov chain Monte Carlo procedure by taking conditional draws of parameters according to the procedure described below, and iterating until convergence. We take explicit draws of the unobserved segment s_i that user i is in through data augmentation. The notation is as follows: $g(\cdot)$ denotes the p.m.f. of the Poisson distribution; $f(\cdot)$ denotes a generic p.d.f.; $\mathbb{I}\{\cdot\}$ is the indicator function; the parameters are as defined in the main text. When we do not have a closed-form expression for the posterior probability, we take draws using a Metropolis-Hastings procedure with random walk.

- Draw $\beta^{\lambda}_{k,s}$ for keyword k and segment s:

$$f(\beta^{\lambda}_{k,s}|\bar{\beta}^{\lambda}_{s},(\sigma^{\lambda}_{s})^{2},\beta^{p}_{k,s},\beta^{\lambda}_{\text{POP}},\beta^{\lambda}_{X},\beta^{\lambda}_{Z},\beta^{p}_{\text{POP}},\beta^{p}_{X},\beta^{p}_{Z},\{s_i\}_{i=1,\ldots,I})$$
$$\propto\left(\prod_{i=1}^{I}\left(g\left(y^{O}_{ki};\lambda_{ki,s}(1-p_{ki,s})\right)\cdot g\left(y^{S}_{ki};\lambda_{ki,s}p_{ki,s}\right)\right)^{\mathbb{I}\{i\text{ searches }k\}\cdot\mathbb{I}\{s_i=s\}}\right)\varphi\left(\beta^{\lambda}_{k,s};\bar{\beta}^{\lambda}_{s},(\sigma^{\lambda}_{s})^{2}\right),$$

  where $\varphi(\cdot)$ is the p.d.f. of the Normal distribution.

- Draw $\beta^{p}_{k,s}$ for keyword k and segment s:

$$f(\beta^{p}_{k,s}|\bar{\beta}^{p}_{s},(\sigma^{p}_{s})^{2},\beta^{\lambda}_{k,s},\beta^{\lambda}_{\text{POP}},\beta^{\lambda}_{X},\beta^{\lambda}_{Z},\beta^{p}_{\text{POP}},\beta^{p}_{X},\beta^{p}_{Z},\{s_i\}_{i=1,\ldots,I})$$
$$\propto\left(\prod_{i=1}^{I}\left(g\left(y^{O}_{ki};\lambda_{ki,s}(1-p_{ki,s})\right)\cdot g\left(y^{S}_{ki};\lambda_{ki,s}p_{ki,s}\right)\right)^{\mathbb{I}\{i\text{ searches }k\}\cdot\mathbb{I}\{s_i=s\}}\right)\varphi\left(\beta^{p}_{k,s};\bar{\beta}^{p}_{s},(\sigma^{p}_{s})^{2}\right).$$

- Draw $\bar{\beta}^{\lambda}_{s}$ for segment s:

$$\bar{\beta}^{\lambda}_{s}|\{\beta^{\lambda}_{k,s}\}_{k=1,\ldots,K},(\sigma^{\lambda}_{s})^{2}\sim\text{Normal}\left(\left(\frac{1}{\sigma_0^2}+\frac{K}{(\sigma^{\lambda}_{s})^{2}}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{k=1}^{K}\beta^{\lambda}_{k,s}}{(\sigma^{\lambda}_{s})^{2}}\right),\left(\frac{1}{\sigma_0^2}+\frac{K}{(\sigma^{\lambda}_{s})^{2}}\right)^{-1}\right).$$

  We choose diffuse conjugate Normal prior for $\bar{\beta}^{\lambda}_{s}$.

- Draw $\bar{\beta}^{p}_{s}$ for segment s:

$$\bar{\beta}^{p}_{s}|\{\beta^{p}_{k,s}\}_{k=1,\ldots,K},(\sigma^{p}_{s})^{2}\sim\text{Normal}\left(\left(\frac{1}{\sigma_0^2}+\frac{K}{(\sigma^{p}_{s})^{2}}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{k=1}^{K}\beta^{p}_{k,s}}{(\sigma^{p}_{s})^{2}}\right),\left(\frac{1}{\sigma_0^2}+\frac{K}{(\sigma^{p}_{s})^{2}}\right)^{-1}\right).$$

  We choose diffuse conjugate Normal prior for $\bar{\beta}^{p}_{s}$.

- Draw $\left(\sigma^{\lambda}_{s}\right)^{2}$ for segment s:

$$(\sigma^{\lambda}_{s})^{2}|\{\beta^{\lambda}_{k,s}\}_{k=1,\ldots,K},\bar{\beta}^{\lambda}_{s}\sim\text{Inv-Gamma}\left(\nu_0+\frac{K}{2},s_0+\sum_{k=1}^{K}\left(\beta^{\lambda}_{k,s}-\bar{\beta}^{\lambda}_{s}\right)^{2}\right).$$

We choose diffuse conjugate Inverse-Gamma prior for $\left(\sigma_s^{\lambda}\right)^2$.

- Draw $\left(\sigma_s^{p}\right)^2$ for segment s:

$$\left(\sigma_s^{p}\right)^2 | \{\beta_{k,s}^{p}\}_{k=1,\ldots,K}, \bar{\beta}_s^{p} \sim \text{Inv-Gamma}\left(\nu_0 + \frac{K}{2}, s_0 + \sum_{k=1}^{K}\left(\beta_{k,s}^{p} - \bar{\beta}_s^{p}\right)^2\right).$$

We choose diffuse conjugate Inverse-Gamma prior for $\left(\sigma_s^{p}\right)^2$.

- Draw $\beta_{\text{POP}}^{\lambda}$, β_X^{λ}, and β_Z^{λ}:

$$\begin{aligned} &f(\beta_{\text{POP}}^{\lambda}, \beta_X^{\lambda}, \beta_Z^{\lambda} | \{\beta_{k,s}^{\lambda}, \beta_{k,s}^{p}\}_{k=1,\ldots,K, s=1,\ldots,S}, \beta_{\text{POP}}^{p}, \beta_X^{p}, \beta_Z^{p}, \{s_i\}_{i=1,\ldots,I}) \\ &\propto \left(\prod_{i=1}^{I}\left(g\left(y_{ki}^{O}; \lambda_{ki,s}(1-p_{ki,s})\right) \cdot g\left(y_{ki}^{S}; \lambda_{ki,s}p_{ki,s}\right)\right)\right) f\left(\beta_{\text{POP}}^{\lambda}\right) f\left(\beta_X^{\lambda}\right) f\left(\beta_Z^{\lambda}\right). \end{aligned}$$

We choose diffuse priors for $\beta_{\text{POP}}^{\lambda}$, β_X^{λ}, and β_Z^{λ}.

- Draw β_{POP}^{p}, β_X^{p}, and β_Z^{p}:

$$\begin{aligned} &f(\beta_{\text{POP}}^{p}, \beta_X^{p}, \beta_Z^{p} | \{\beta_{k,s}^{\lambda}, \beta_{k,s}^{p}\}_{k=1,\ldots,K, s=1,\ldots,S}, \beta_{\text{POP}}^{\lambda}, \beta_X^{\lambda}, \beta_Z^{\lambda}, \{s_i\}_{i=1,\ldots,I}) \\ &\propto \left(\prod_{i=1}^{I}\left(g\left(y_{ki}^{O}; \lambda_{ki,s}(1-p_{ki,s})\right) \cdot g\left(y_{ki}^{S}; \lambda_{ki,s}p_{ki,s}\right)\right)\right) f\left(\beta_{\text{POP}}^{p}\right) f\left(\beta_X^{p}\right) f\left(\beta_Z^{p}\right). \end{aligned}$$

We choose diffuse priors for β_{POP}^{p}, β_X^{p}, and β_Z^{p}.

- Draw s_i for search i for keyword k:

$$\begin{aligned} &f(s_i | \{\beta_{k,s}^{\lambda}, \beta_{k,s}^{p}\}_{k=1,\ldots,K, s=1,\ldots,S}, \alpha_1, \beta_{\text{POP}}^{\lambda}, \beta_X^{\lambda}, \beta_Z^{\lambda}, \beta_{\text{POP}}^{p}, \beta_X^{p}, \beta_Z^{p}, \pi_k) \\ &\propto \Pr(i \text{ searches } k) \cdot g\left(y_{ki}^{O}; \lambda_{ki,s}(1-p_{ki,s})\right) \cdot g\left(y_{ki}^{S}; \lambda_{ki,s}p_{ki,s}\right) h(s_i; \pi_k), \end{aligned}$$

where $h(\cdot)$ is the p.m.f. of the Categorical distribution. This is the data augmentation step.

- Draw π_k for keyword k :

$$\pi_k | \{s_i\}_{i=1,\ldots,I} \sim \text{Dirichlet}\left(c_1, \ldots, c_S\right),$$

where $c_s = c_0 + \sum_{i=1}^{I}(\mathbb{I}\{i \text{ searches } k\} \cdot \mathbb{I}\{s_i = s\})$ and $c_0 = 1$. That is, we choose diffuse conjugate Dirichlet prior for π_k.