EXAMINING THE IMPACT OF KEYWORD AMBIGUITY ON SEARCH ADVERTISING PERFORMANCE: A TOPIC MODEL APPROACH

Jing Gong  
Department of Management Information Systems, Fox School of Business, Temple University, Philadelphia, PA 19103 U.S.A. {gong@temple.edu} 

Vibhanshu Abhishek  
Paul Merage School of Business, University of California – Irvine, Irvine, CA 92697 U.S.A. {vibs@uci.edu} 

Beibei Li  
Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213 U.S.A. {beibeili@andrew.cmu.edu} 

Appendix A  
Summary of Empirical Studies on Sponsored Search Advertising

<table>
<thead>
<tr>
<th>Paper</th>
<th>Goal</th>
<th>Data Source</th>
<th>Industry</th>
<th>Level of Detail</th>
<th>Number of Keywords Examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarwal et al. (2011)</td>
<td>Impact of position on click-through and conversion</td>
<td>Advertiser</td>
<td>Pet products</td>
<td>Aggregate</td>
<td>68</td>
</tr>
<tr>
<td>Agarwal et al. (2015)</td>
<td>Impact of organic competition on click-through and conversion</td>
<td>Advertiser</td>
<td>Pet products</td>
<td>Aggregate</td>
<td>36</td>
</tr>
<tr>
<td>Chan et al. (2011)</td>
<td>Measuring the value of customers acquired from sponsored search</td>
<td>Advertiser</td>
<td>Lab supplies</td>
<td>Individual</td>
<td>90-208</td>
</tr>
<tr>
<td>Chan and Park (2015)</td>
<td>Advertiser valuation of consumer search activities</td>
<td>Search engine</td>
<td>Sporting goods</td>
<td>Individual</td>
<td>1</td>
</tr>
<tr>
<td>Ghose and Yang (2009)</td>
<td>Impact of keyword attributes on click-through and conversion</td>
<td>Advertiser</td>
<td>Retail</td>
<td>Aggregate</td>
<td>1,878</td>
</tr>
<tr>
<td>Goldfarb and Tucker (2011)</td>
<td>Online and offline advertising channel substitution</td>
<td>Advertiser</td>
<td>Legal service</td>
<td>Aggregate</td>
<td>139</td>
</tr>
<tr>
<td>Jerath et al. (2014)</td>
<td>Impact of keyword popularity on click performance</td>
<td>Search engine</td>
<td></td>
<td>Individual</td>
<td>1,200</td>
</tr>
<tr>
<td>Jeziorski and Segal (2015)</td>
<td>Quantifying rational user experience and externalities among ads</td>
<td>Search engine</td>
<td></td>
<td>Individual</td>
<td>4</td>
</tr>
<tr>
<td>Rutz et al.</td>
<td>Impact of ad position on</td>
<td>Advertiser</td>
<td>Hotel</td>
<td>Aggregate</td>
<td>301</td>
</tr>
</tbody>
</table>
### Appendix B

**Latent Dirichlet Allocation**

The most widely used topic model is the latent Dirichlet allocation model (LDA; Blei et al. 2003), which is a hierarchical Bayesian model that describes a generative process of document creation. The goal of LDA is to infer topics as latent variables from the observed distribution of words in each document. In particular, a topic is defined as a multinomial distribution over a vocabulary of words, a document is a collection of words drawn from one or more topics, and a corpus is the set of all documents. Based on our discussion on corpus construction, we construct a document for each keyword that best reflects the contextual information of the keyword. We now discuss how we use LDA to infer the topics from the corpus of documents.

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

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

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

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

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

Based on the generative process described above, we use a Markov chain Monte Carlo (MCMC) algorithm to estimate $\phi$ and $\theta$. Specifically, we use a collapsed Gibbs sampler to sequentially sample the topic of each word token in the corpus conditional on the current topic assignments of all other word tokens (for details, see Griffiths and Steyvers 2004). We run a collapsed Gibbs sampler using MALLET (McCallum 2002) with 2,000 iterations.
Appendix C

Topic Distribution of Sample Keywords

Figure C1 illustrates the topic distribution of some sample keywords. In Figure C1, topics are labeled on the horizontal axis, and keywords are labeled on the vertical axis. The size of each bubble indicates a posterior topic probability, with larger bubbles representing higher probabilities. For example, the top-left bubble represents the posterior probability that the keyword “judges gavels” belongs to the topic “music,” which is much smaller than the posterior probability that “judges gavels” belongs to “government,” represented by the eighth bubble on the first row. Meanwhile, the keyword “marriage records” has a much larger posterior probability of belonging to the topic “government,” which suggests “marriage records” is most likely related to government affairs rather than other topics.
Appendix D

Extracting Brand and Location Information

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

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

Appendix E

Extracting Transactional Intent

In this study, we are interested in learning how likely consumers are to engage in a transaction when they search for a keyword. Therefore, we focus on detecting transactional intent from keywords. Some keywords may contain explicit transactional words, such as “cheap hotels” and “cruise deals,” but most keywords don’t contain explicit transactional indicators in the keywords, such as “airline tickets” and “honda parts.” The augmented Google organic search results, on the other hand, provide a better picture in terms of consumer search intent. If the keyword has a transactional intent, the Google organic search results are likely to contain transactional indicators such as “buy,” “discount,” “promotion,” and “check out.” Therefore, we propose to infer transactional intent using the keyword’s corresponding Google organic results. First, we compose a list of transactional words based on Dai et al. (2006) and general knowledge. These transactional words are listed in Table E1. Then, for each search keyword, we count the frequency of transactional words in the corresponding Google organic results. We use \( \text{LOG}_{\text{TRANS}} \), the natural log of the frequency of transactional words, to measure keyword’s transactional intent.

<table>
<thead>
<tr>
<th>Table E1. Transactional Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>advertise</td>
</tr>
<tr>
<td>auction</td>
</tr>
<tr>
<td>bidding</td>
</tr>
<tr>
<td>bill</td>
</tr>
<tr>
<td>book</td>
</tr>
<tr>
<td>buy</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

1As a robustness check, we also use the “n-grams” method for string matching, where we define \( n = 2, 3, \) and 4. We find the final results remain consistent.

2To choose the optimal cut-off distance, we first manually identified whether a smaller number of keywords contain brand names. We then tried different cut-off values and chose the one (e.g., 0.85) that minimizes classification errors on the small keyword set.

3Similar to the process of identifying brand names, we chose the optimal cut-off distance based on a smaller set of keywords to minimize the classification error.
Appendix F

Correlation among Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>TOPIC_ENTROPY</th>
<th>NUM_WORDS</th>
<th>BRAND</th>
<th>LOCATION</th>
<th>LOG_TRANS</th>
<th>LOG_IMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPIC_ENTROPY</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUM_WORDS</td>
<td>-0.38</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAND</td>
<td>-0.03</td>
<td>0.06</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.00</td>
<td>0.16</td>
<td>0.04</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>0.07</td>
<td>-0.04</td>
<td>0.22</td>
<td>0.01</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>LOG_IMP</td>
<td>-0.03</td>
<td>-0.17</td>
<td>0.19</td>
<td>-0.02</td>
<td>0.08</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Appendix G

The Gibbs Sampling Procedure

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

For simplification, we assume that

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

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

$$u^k = (u_{0,k}, u_{1,k})'$$

$$\Delta^k = (\Delta^0, \Delta^1)'$$

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

The hierarchical Bayesian model can be written in the following hierarchical form:
We assume the following prior specifications:

\[
\begin{align*}
\beta_2 & \sim N(0, \nu^{\beta_2}) \\
\phi_2 & \sim N(0, \nu^{\phi_2}) \\
\text{vecr}(\Delta^\theta) & \sim MVN(\bar{\Delta}^\theta, A^\theta) \\
\text{vecr}(\Delta^\phi) & \sim MVN(\bar{\Delta}^\phi, A^\phi) \\
v^T & \sim IG(m, n) \\
\sigma_{12}^2 & \sim N(r_{12}, b_{12}) \\
\sigma^2 & \sim IG(v_{12}, c_{12}) \\
\psi & \sim IW(v^\psi, V^\psi) \\
\Omega^\beta & \sim IW(v^{\Omega^\beta}, V^{\Omega^\beta}) \\
\Omega^\phi & \sim IW(v^{\Omega^\phi}, V^{\Omega^\phi})
\end{align*}
\]

We describe the Gibbs sampling procedure below.

**Step 1. Draw** \(t_i|\widehat{\theta}_{k_i}\) **for each impression** \(i\).

**Step 2. Draw** \(U_{lat}\) **for each observation.**

We can draw \(U_{lat}\) from the following posterior distribution:

\[
U_{lat} \sim TN(\mu_{lat}, \sigma_{1|2})
\]

where \(TN\) denotes the truncated normal distribution, and \(U_{lat}\) is truncated above zero if \(Click_{ia} = 1\), and below zero if \(Click_{ia} = 0\). Let \(\bar{U}_{lat} = \beta_{0,ki} + \beta_{1,ki} POS_{ia} + \beta_{2}NUM_{AD_{i}} + \tau_{ia}, \epsilon_{ia} = POS_{ia} - \phi_{0,ki} - \phi_{1,ki} POS_{a,-ki} - \phi_{2}NUM_{AD_{i}}\), then

\[
\begin{align*}
\mu_{lat} & = \bar{U}_{lat} + \frac{\sigma_{12}}{\sigma^2 + \sigma_{12}} \epsilon_{ia} \\
\sigma_{1|2} & = 1 - \frac{\sigma_{12}}{\sigma^2 + \sigma_{12}} = \frac{\sigma^2}{\sigma^2 + \sigma_{12}}
\end{align*}
\]

**Step 3. Draw** \(\Gamma_k = (\chi^k, \phi^k)\) **for each keyword** \(k\).

For each keyword \(k\), let \(N_k\) be the number of observations such that \(k_i = k\). Let \(\Gamma_k = (\chi^k, \phi^k)\), \(z_{1ia} = (1, POS_{ia})\), \(z_{2ia} = (1, POS_{a,-ki})\), \(y_{1ia} = U_{lat} - z'_{1ia} \gamma - \beta_{2}NUM_{AD_{i}} - \tau_{ia}, y_{2ia} = POS_{ia} - \phi_{2}NUM_{AD_{i}}\). Then

\[
\begin{align*}
y_{1ia} & = z'_{1ia} \chi^k + \eta_{ia} \\
y_{2ia} & = z'_{2ia} \phi^k + \epsilon_{ia}
\end{align*}
\]

where \((\eta_{ia}, \epsilon_{ia})\) \(\sim MVN(0, \Lambda)\). We can write it in matrix version as
\[ y_{1k} = Z'_{1k} \alpha^k + \eta_k \\
 y_{2k} = Z'_{2k} \phi^k + \epsilon_k \]

or more compactly, as

\[ Y = Z' \Gamma_k + E_k \]

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

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

With prior distribution \( \Gamma_k \sim MVN(\bar{\Gamma}_k, \psi_0) \), where \( \bar{\Gamma}_k = \{I_N \otimes X_k\} \begin{pmatrix} \delta^{\beta}_{\delta^\beta} \\ \delta^{\phi}_{\delta^\phi} \end{pmatrix} \) and \( \psi_0 = \begin{pmatrix} \Omega^\beta_{\delta^\beta} & 0 \\ 0 & \Omega^\phi_{\delta^\phi} \end{pmatrix} \), we can draw \( \Gamma_k \) from the following posterior distribution:

\[ \Gamma_k|\text{all other parameters} \sim MVN(\tilde{\Gamma}_k, \psi) \]

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

**Step 4. Draw \( \gamma_t \) for each topic \( t. \)**

For each topic \( t \), let \( N_t \) be the number of observations with \( t_i = t \). Let \( Z_{ia} = (1, POS_{ia})' \), \( \epsilon_{ia} = POS_{ia} - \phi_{0, ki} - \phi_{1, ki} - \phi_{2, NUM AD_i} \),

\[ U_{iat} = U_{lat} - Z_{ia} \alpha^k - \beta_2 NUM AD_i - \tau_a - \frac{\alpha_{22}}{\sigma^2 + \sigma_{t2}^2} \epsilon_{ia} \]

and \( \sigma_{12} = \frac{\sigma^2}{\sigma^2 + \sigma_{t2}^2} \). Then \( U_{lat} \sim N(Z'_{ia} Y_t, \sigma_{11}) \). We can write it in matrix version as

\[ U_{1t} \sim MVN(Z' Y_t, \sigma_{11} I_{N_t}) \]

where \( U_{1t} : N_t \times 1 \) includes all \( U_{lat} \) such that \( t_i = t \), and \( Z_t \) is a \( N_t \times 2 \) matrix. With prior \( \gamma_t \sim MVN(0, \Psi) \), we then draw \( \gamma_t \) from the following posterior distribution:

\[ \gamma_t|\text{all other parameters} \sim MVN(\tilde{\gamma}_t, \Phi) \]

where \( \Phi = [\Psi^{-1} + (Z' Y_t Z_t)^{-1}]^{-1} \) and \( \tilde{\gamma}_t = \Phi(Z' Y_t U_{1t}) \).

**Step 5. Draw \( \tau_a \) for each ad \( a. \)**

For each ad \( a \), let \( n_a \) be the number of observations. We define \( \epsilon_{iat} = POS_{ia} - \phi_{0, ki} - \phi_{1, ki} - \phi_{2, NUM AD_i} \),

\[ U_{iat} = U_{lat} - (\beta_{0, ki} + \beta_{1, ki} POS_{ia}) - \beta_2 NUM AD_i - \frac{\alpha_{22}}{\sigma^2 + \sigma_{t2}^2} \epsilon_{ia} \]

and \( \sigma_{12} = \frac{\sigma^2}{\sigma^2 + \sigma_{t2}^2} \). Then \( U_{lat} \sim N(\tau_a, \sigma_{11}) \). With prior \( \tau_a \sim N(0, \nu^2) \), the posterior distribution of \( \tau_a \) is

\[ \tau_a|\text{all other parameters} \sim MVN(\tilde{\tau}_a, \tilde{\nu}^2) \]

where \( \tilde{\tau}_a = \frac{n_a \nu^2 U_{2t}^2}{n_a \nu^2 + \sigma_{11}} \) and \( \tilde{\nu}^2 = \frac{\sigma_{11} \nu^2}{n_a \nu^2 + \sigma_{11}} \).
Step 6. Draw $\beta_2$.

Let $\epsilon_{iat} = POS_{iat} - \phi_{0,k_i} - \phi_{1,k_i} - \phi_2 NUM_{AD_i}$, $U_{iat} = \epsilon_{iat} - \beta_0 k_i t + \beta_1 k_i t POS_{iat} - \beta_2 NUM_{AD_i} - \frac{\sigma_{\epsilon}}{\sigma_{\epsilon} + \sigma_2} \epsilon_{iat}$. $X_i = NUM_{AD_i}$, then $U_{iat} \sim N(\beta_2 X_i, \sigma_{\epsilon}^2)$. With prior $\beta_2 \sim N(0, \nu_{\beta_2})$, the posterior distribution of $\beta_2$ is

$$\beta_2 | \text{all other parameters} \sim N(\bar{\beta}_2, \nu_{\beta_2})$$

where $\nu_{\beta_2} = \left[ \sigma_{\epsilon}^{-1} X' X + (\nu_{\beta_2})^{-1} \right]^{-1}$, and $\bar{\beta}_2 = \nu_{\beta_2} \left[ \sigma_{\epsilon}^{-1} X' U \right]$.

Step 7. Draw $\phi_2$.

Let $\tilde{\eta}_{iat} = U_{iat} - (\beta_0 k_i t + \beta_1 k_i t POS_{iat} - \beta_2 NUM_{AD_i} - \sigma_{\epsilon}^2 \tilde{\epsilon}_{iat})$, $\omega_{iat} = POS_{iat} - (\phi_{0,k_i} + \phi_1 k_i t POS_{iat} - \sigma_{\epsilon}^2 \tilde{\epsilon}_{iat})$. $X_i = NUM_{AD_i}$, then $\omega_{iat} \sim N(\phi_2 X_i, \nu_{\phi_2})$. With prior $\phi_2 \sim N(0, \nu_{\phi_2})$, the posterior distribution of $\phi_2$ is

$$\phi_2 | \text{all other parameters} \sim N(\bar{\phi}_2, \nu_{\phi_2})$$

where $\nu_{\phi_2} = \left[ \nu_2^{-1} X' X + (\nu_{\phi_2})^{-1} \right]^{-1}$, and $\bar{\phi}_2 = \nu_{\phi_2} \left[ \nu_2^{-1} X' \omega \right]$.

Step 8. Draw $\Delta^\beta$.

Let $K$ be the number of keywords. With $\Delta^\beta = \left( \Delta^\beta_{11}, \ldots, \Delta^\beta_{1r}, \ldots, \Delta^\beta_{rr} \right)$, we have $\Delta^\beta = \text{vecr}(\Delta^\beta) = \left( \Delta^\beta_{11}, \ldots, \Delta^\beta_{rr}, \Delta^\beta_{21}, \ldots, \Delta^\beta_{2r} \right)^T$. Therefore, $\chi^k = (X_k' 0)

where $\chi^k = (X_1, \ldots, X_k)^T$, $X = (X_1, \ldots, X_k)$, and $E \sim MVN(0, \Omega^\beta \otimes I_K)$. More compactly, we can write

$$\chi = (I_k \otimes X) \Delta^\beta + E$$

With prior $\Delta^\beta \sim MVN(\tilde{\Delta}^\beta, \Lambda^\beta)$, we can draw $\Delta^\beta$ from the following posterior distribution:

$$\Delta^\beta | \text{all other parameters} \sim MVN(\tilde{\Delta}^\beta, \Lambda^\beta)$$

where $\Lambda^\beta = (I_k \otimes X)^T \left( \Omega^\beta^{-1} \otimes I_k \right) (I_k \otimes X) + (\Lambda^\beta)^{-1}$ and $\tilde{\Delta}^\beta = \Lambda^\beta \left( (I_k \otimes X)^T \left( \Omega^\beta^{-1} \otimes I_k \right) X + (\Lambda^\beta)^{-1} \tilde{\Delta}^\beta \right)$.

Step 9. Draw $\Delta^\phi$.

Similar to the previous step, with prior $\Delta^\phi \sim MVN(\tilde{\Delta}^\phi, \Lambda^\phi)$, we can draw $\Delta^\phi$ from the following posterior distribution:

$$\Delta^\phi | \text{all other parameters} \sim MVN(\tilde{\Delta}^\phi, \Lambda^\phi)$$
where $\tilde{A} = \left( I_2 \otimes X \right) \left( \left( \Omega^\phi \right)^{-1} \otimes I_K \right) \left( I_2 \otimes X \right) + \left( A^\phi \right)^{-1}$ and $\tilde{A} = \tilde{A} \left[ I_2 \otimes X \right] \left( \left( \Omega^\phi \right)^{-1} \otimes I_K \right) \chi + \left( A^\phi \right)^{-1} \tilde{A}$.  

**Step 10. Draw $v^\tau$.**

With prior $v^\tau \sim IG(m, n)$, we can draw $v^\tau$ from its posterior distribution $IG \left( m + A, n + \sum_{i=1}^T \gamma_i \right)$, where $A$ is the total number of unique ads.

**Step 11. Draw $\sigma_{12}$.**

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

**Step 12. Draw $\sigma^2$.**

With prior $\sigma^2 \sim IG(v_0, c_0)$, we can draw $\sigma^2$ from its posterior distribution $IG \left( v_0 + \frac{N}{2}, c_0 + \frac{1}{2} \right)$, where $N$ is the number of observations.

**Step 13. Draw $\Omega^\beta$.**

With prior $\Omega^\beta \sim IW \left( v^\beta, V^\beta \right)$, we can draw $\Omega^\beta$ from its posterior distribution $IW \left( v^\beta + K, V^\beta + \sum_{k=1}^K (\chi^k - \Delta^\beta X_k) \left( \chi^k - \Delta^\beta X_k \right)^\top \right)$, where $K$ is the number of keywords.

**Step 14. Draw $\Omega^\phi$.**

With prior $\Omega^\phi \sim IW \left( v^\phi, V^\phi \right)$, we can draw $\Omega^\phi$ from its posterior distribution $IW \left( v^\phi + K, V^\phi + \sum_{k=1}^K (\phi^k - \Delta^\phi X_k) \left( \phi^k - \Delta^\phi X_k \right)^\top \right)$, where $K$ is the number of keywords.

**Step 15. Draw $\Psi$.**

With prior $\Psi \sim IW \left( v^\Psi, V^\Psi \right)$, we can draw $\Psi$ from its posterior distribution $IW \left( v^\Psi + T, V^\Psi + \sum_{t=1}^T \gamma_t \gamma_t^\top \right)$, where $T$ is the number of topics.
Appendix H

Analysis on Time Before First Click

Although we do not directly observe consumers’ click behavior on organic search results, we are able to infer the time on the organic links by observing when a consumer starts a search session by entering a keyword, and when the consumer clicks on the first ad. Therefore, we focus our analysis on the subset of consumers who have clicked on at least one sponsored ad. We denote $DURATION$ as the time between the consumer starting a search session and making the first click, and we use $DURATION$ as a proxy for the time spent on the organic listing. We run a linear regression of $DURATION$ on the keyword attributes of interests. As shown in Table H1, higher keyword ambiguity is associated with less time spent on organic search results, while more precise keywords tend to attract more attention on organic search results. This result provides partial evidence that a more ambiguous keyword may reduce the attractiveness of organic search results, and consumers may turn to sponsored ads for finding an alternative that meets their needs.

| Table H1. Regression Results: Time Before First Click |
|-----------------|-----------------|
| Variable        | Estimates        |
| Intercept       | 56.925*** (0.675)|
| $TOPIC\_ENTROPY$| -4.643*** (0.232)|
| $NUM\_WORDS$    | 0.335** (0.155)  |
| $BRAND$         | -2.068*** (0.210)|
| $LOCATION$      | -1.200*** (0.301)|
| $LOG\_TRANS$   | -2.895*** (0.090)|
| $LOG\_IMP$      | -2.674*** (0.057)|
| Observations    | 551,239          |

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

Appendix I

Boxplots of Topic-Specific Effects

To present the heterogeneous impact on CTR across topics, we overlay boxplots of the topic specific intercepts (i.e., $\gamma_0$) in Figure I1 and the topic specific effects of position (i.e., $\gamma_1$) in Figure I2. Because we have mean-centered all keyword characteristics when estimating the hierarchical Bayesian model, $\gamma_0$ and $\gamma_1$ can be interpreted as estimates of $\beta_0, k$ and $\beta_1, k$ for a typical keyword in topic $t$ of which the covariates are set to mean values. As we can see from Figure I1, the means of the posterior distribution of the topic specific intercepts are highest for topics “travel” and “health,” suggesting that consumers who are interested in those topics may be more likely to click on ads at top positions. In contrast, for topics “sport” and “adult,” CTR is lower at top positions.

As we can see from Figure I2, the means of the posterior distribution of the topic specific effects of position are highest for topics “home” and “documents,” suggesting that consumers who are interested in those topics may be more likely to click on ads at lower positions. In contrast, for topics “music” and “clothing,” CTR decreases faster with position.
Figure I1. Boxplots of Topic-Specific Intercepts ($\gamma_{0,t}$)
Appendix J

Number of Organic Results for Corpus Construction

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

Comparing topic entropy. In Table J1, we present the summary statistics for the computed topic entropy of the full data set (12,790 keywords) based on different numbers of organic search results. The high correlations among entropy values derived based on different numbers of organic search results suggest that entropy values seem to be fairly robust to the number of organic search results used to construct the corpus for topic modeling.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 50</td>
<td>1.60</td>
<td>0.45</td>
<td>0.34</td>
<td>Top 50</td>
</tr>
<tr>
<td></td>
<td>Top 60</td>
<td>1.97</td>
<td>0.41</td>
<td>0.44</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>Top 80</td>
<td>1.99</td>
<td>0.40</td>
<td>0.44</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>Top 100</td>
<td>2.00</td>
<td>0.40</td>
<td>0.44</td>
<td>3.00</td>
</tr>
</tbody>
</table>
Comparing empirical results. We have further reestimated the hierarchical Bayesian model using the entropy values and topic probabilities we now obtained based on different numbers of organic search results. We present the main results for CTR in Table J2. As can be seen, the estimation results are fairly consistent across different columns, suggesting that our main results are robust to the number of organic search results used to cover the topics related to each keyword.

### Table J2. Estimation Results for CTR Based on Different Number of Google Organic Search Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Top 50</th>
<th>Top 60</th>
<th>Top 80</th>
<th>Top 100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.308*** (0.122)</td>
<td>-2.336*** (0.126)</td>
<td>-2.316*** (0.108)</td>
<td>-2.318*** (0.115)</td>
</tr>
<tr>
<td>TOPIC_ENTROPY</td>
<td>0.192*** (0.069)</td>
<td>0.223*** (0.074)</td>
<td>0.184** (0.076)</td>
<td>0.165** (0.072)</td>
</tr>
<tr>
<td>NUM_W ORDS</td>
<td>0.040 (0.045)</td>
<td>0.049 (0.045)</td>
<td>0.038 (0.046)</td>
<td>0.033 (0.045)</td>
</tr>
<tr>
<td>BRAND</td>
<td>0.033 (0.064)</td>
<td>0.042 (0.064)</td>
<td>0.040 (0.064)</td>
<td>0.040 (0.064)</td>
</tr>
<tr>
<td>LOCATION</td>
<td>-0.208** (0.105)</td>
<td>-0.217** (0.100)</td>
<td>-0.212** (0.098)</td>
<td>-0.207** (0.098)</td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>0.147*** (0.029)</td>
<td>0.150*** (0.030)</td>
<td>0.152*** (0.029)</td>
<td>0.153*** (0.029)</td>
</tr>
<tr>
<td>LOG_IMP</td>
<td>-0.010 (0.019)</td>
<td>-0.004 (0.020)</td>
<td>-0.004 (0.019)</td>
<td>-0.005 (0.019)</td>
</tr>
<tr>
<td><strong>POS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.726*** (0.045)</td>
<td>-0.749*** (0.066)</td>
<td>-0.727*** (0.059)</td>
<td>-0.724*** (0.055)</td>
</tr>
<tr>
<td>TOPIC_ENTROPY</td>
<td>-0.134*** (0.049)</td>
<td>-0.115** (0.048)</td>
<td>-0.114** (0.048)</td>
<td>-0.113** (0.049)</td>
</tr>
<tr>
<td>NUM_W ORDS</td>
<td>-0.035 (0.032)</td>
<td>-0.035 (0.032)</td>
<td>-0.034 (0.030)</td>
<td>-0.034 (0.031)</td>
</tr>
<tr>
<td>BRAND</td>
<td>-0.050 (0.045)</td>
<td>-0.056 (0.047)</td>
<td>-0.051 (0.044)</td>
<td>-0.047 (0.045)</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0.070 (0.064)</td>
<td>0.075 (0.065)</td>
<td>0.072 (0.063)</td>
<td>0.075 (0.063)</td>
</tr>
<tr>
<td>LOG_TRANS</td>
<td>-0.014 (0.019)</td>
<td>-0.013 (0.019)</td>
<td>-0.015 (0.019)</td>
<td>-0.014 (0.019)</td>
</tr>
<tr>
<td>LOG_IMP</td>
<td>-0.018 (0.013)</td>
<td>-0.019 (0.013)</td>
<td>-0.021* (0.013)</td>
<td>-0.021 (0.013)</td>
</tr>
<tr>
<td><strong>NUM_ADS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.166*** (0.017)</td>
<td>0.168*** (0.018)</td>
<td>0.167*** (0.015)</td>
<td>0.165*** (0.016)</td>
</tr>
</tbody>
</table>

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

### References


