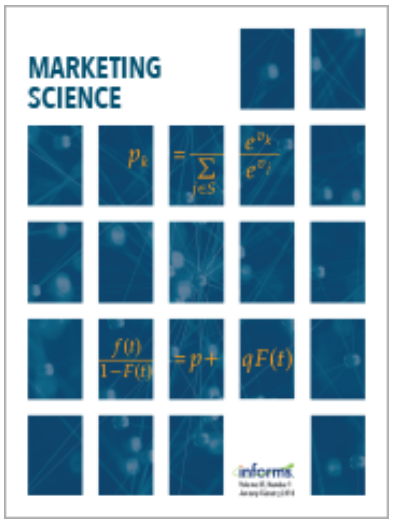


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting

Michael Trusov, Liye Ma, Zainab Jamal

To cite this article:
Michael Trusov, Liye Ma, Zainab Jamal (2016) Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting. Marketing Science 35(3):405-426. https://doi.org/10.1287/mksc.2015.0956

MARKETING SCIENCE
Vol. 35, No. 3, May–June 2016, pp. 405–426
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2015.0956

Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting

Michael Trusov, Liye Ma
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742
{mtrusov@rhsmith.umd.edu, liyema@rhsmith.umd.edu}

Zainab Jamal
HP Inc., Palo Alto, California 94034, zainab.jamal@hp.com

User profile is a summary of a consumer's interests and preferences revealed through the consumer's online activity. It is a fundamental component of numerous applications in digital marketing. McKinsey & Company view online user profiling as one of the promising opportunities companies should take advantage of to unlock "big data's" potential. This paper proposes a modeling approach that uncovers individual user profiles from online surfing data and allows online businesses to make profile predictions when limited information is available. The approach is easily parallelized and scales well for processing massive records of user online activity. We demonstrate application of our approach to customer-base analysis and display advertising. Our empirical analysis uncovers easy-to-interpret behavior profiles and describes the distribution of such profiles. Furthermore, it reveals that even for information-rich online firms profile inference that is based solely on their internal data may produce biased results. We find that although search engines cover smaller portions of consumer Web visits than major advertising networks, their data is of higher quality. Thus, even with the smaller information set, search engines can effectively recover consumer behavioral profiles. We also show that temporal limitations imposed on individual-level tracking abilities are likely to have a differential impact across major online businesses, and that our approach is particularly effective for temporally limited data. Using economic simulation we demonstrate potential gains the proposed model may offer a firm if used in individual-level targeting of display ads.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2015.0956.

Keywords: big data; user profiling; behavioral targeting; topic models; Internet marketing
History: Received: December 30, 2013; accepted: June 24, 2015; Pradeep Chintagunta, Dominique Hanssens, and John Hauser served as the special issue editors and Daniel Goldstein served as associate editor for this article. Published online in *Articles in Advance* April 28, 2016.

1. Introduction

Consumer information is one of the most important strategic assets of online marketers. There are several components that constitute consumer information, from transaction records to clickstreams to user-generated content. Information availability on each user varies, and different online businesses (e.g., e-commerce sites, search engines, social networking outlets, news portals, advertising networks) have access to different elements of the consumer's online activities that become available to them in the course of interactions with their users.

An e-commerce website, for example, may track a user across multiple visits to its digital storefront, and keep a history of purchases, products viewed on each visit, and abandoned shopping carts. On the other hand, online stores know little about what else their customers do online. By contrast, a popular search engine such as Google tracks user's search activities (i.e., key phrases searched for) as well as which websites are visited immediately after a successful search (i.e., downstream sites). It is unlikely, however, that the search engine would know what the users do on these sites after they leave the search engine.[1] Advertising networks, such as Advertising.com or ValueClick.com, track user activities across multiple sites participating in their network. However, the level of detail about each user's online behavior is limited to the information that network participants have agreed to share with the network. Also, an advertising network's view of the consumer is limited to the information from the sites participating in the network.

Information collected by observing an individual's activity online can be used for different purposes. The most popular application for on-site (or site-centric) data is usability studies and content and navigation optimization, where the firm analyzes how its customers interact with the site and makes adjustments to the site design to improve user experience or goal

[1] This is true unless the site is partnering with the search engine through some explicit or implicit information-sharing programs such as Google Analytics.

conversions, or both. More specialized applications include product recommendation systems (Bodapati 2008, Ghose et al. 2012), dynamic content generation and personalization (Hauser et al. 2009, Urban et al. 2014), and offer customization (Ansari and Mela 2003). Paired with transaction history data, on-site activity records can be used to assess a customer's lifetime value to a firm and churn probability, and enhance responsiveness to promotions and recommendations (Johnson et al. 2004, Moe 2003, Moe and Fader 2004). Search engines mine user records to improve relevance of search results and targeting of their paid search advertising (Rutz and Bucklin 2011, Rutz et al. 2011, Yang and Ghose 2010, Ghose and Yang 2009, Jerath et al. 2014). Advertising networks typically focus on improving responses to advertising (e.g., click-through rate (CTR) or other forms of interaction) as demanded by advertisers.

While the applications of online data vary across different types of online business, *user profiling* is valued universally, as it forms the foundation of these applications. A user's online profile is a summary of a user's interests and preferences revealed through the user's online activity. It is also a cornerstone of behavioral targeting used to tailor advertising and product or service offerings to the user to improve response rates. User profiling, however, is also challenging. Collecting comprehensive information about consumer online activities is a necessary first step for user profiling. Naturally, the best way to build a comprehensive user profile is to collect data on the user's end (e.g., using tracking software installed on a personal Internet device). This is now practiced by a number of marketing research firms, each using a large panel of online users and tracking their activities. Even though technology enables the collection of a vast amount of consumer information, how to extract meaningful consumer interests and behavioral patterns from such data remains largely an open question. Furthermore, most online businesses face the added challenge of profile fragmentation: Although having a complete view on the user is desired, only partial views are available. While it is not feasible for most of these businesses to collect complete user information as the marketing research firms do using their panels, they can usually purchase anonymized data of such panels from third parties. Such anonymized data on a large set of Internet users must then be combined with the internal data available to a specific online firm (i.e., first-party data) to make predictions about missing or incomplete dimensions of the user's profile. Given such situations and challenges, user profiling calls for a solution that can (1) efficiently extract consumer behavioral profiles from a large set of online consumer data, (2) make inferences when only limited information is available, and (3) seamlessly combine internal and external information to make predictions.

Motivated by these important needs, in this paper we develop an original modeling approach that uncovers consumer interests and behavioral patterns from large amounts of online visit information, and allows online firms to make such inferences when only limited information is available to them. Inspired by topic models used in machine learning and natural language processing, our proposed approach represents an individual user profile as a combination of behavioral roles. Intuitively, a user could take on different roles while surfing the Internet, for online shopping, social networking, business, etc. Users taking different roles have distinct propensities to visit websites of different categories. The composition of roles may also vary across users and evolve over time. We demonstrate the application of our modeling approach in two popular areas of Internet marketing: customer-base analysis and display advertising. The ability to handle large data sets and to make inferences based on limited data is vital for these applications. Our proposed approach also offers significant benefits specific to each application, i.e., results interpretability in customer-base analysis and real-time performance in display advertising.

The issue of online user profiling naturally lands itself in the domain of big data. First, user online activity records are vast and require scalable algorithms for processing. With roughly 280 million Internet users in the United States alone (InternetLiveStat.com 2014) online marketers face a challenging problem of making meaningful inferences about individual consumers using digital footprints they leave online. Second, profile inference decisions in some popular applications such as auction-based display advertising must be done in real time.[2] Finally, while the managerial input in such environments still plays an important role, profile prediction in many applications needs to run in a fully automated fashion.

The main contribution of this paper is the development of an effective modeling approach to address the issue of user profiling in digital marketing. Our modeling approach is grounded in the Correlated Topic Model (CTM) (Blei and Lafferty 2007). Topic models were originally developed in the domain of document classification, but their conceptual framework is a good fit for user profiling. Our model makes major extensions from the original CTM model to incorporate visitation intensity, heterogeneity, and dynamics. Our approach shows strong profile recovery performance and is highly scalable to large data sets. We demonstrate how

[2] For example, to minimize page-load time and improve user experience, DoubleClick ad exchange expects to receive each bid submission request within 100 milliseconds after the advertiser is notified about impression opportunity.

the proposed method can be set to run on multiple CPUs. Having light overhead, our algorithm is easily parallelized and improves performance almost linearly in the number of CPUs. An equally important benefit of our proposed approach is result interpretability. We demonstrate that our modeling approach uncovers user profiles that are easily interpretable and intuitive, while taking into account observed and unobserved heterogeneities. This important aspect of our proposed approach makes it appealing to marketing practitioners who seek to understand behavioral composition of their site visitors. Using economic simulation, we illustrate potential gains the proposed model may offer to a firm if used in individual-level targeting of display ads. As a substantive contribution, we show that even for highly visible digital companies (e.g., Google), profile inference based solely on the first-party data may produce biased results (i.e., over- or underrepresenting certain behavioral aspects of a user profile). The method developed in this paper helps to correct for these biases. Finally, we argue that the proposed approach may be of value in a number of applications beyond the marketing field such as policy making and consumer privacy regulations.

The paper is organized as follows. In §§2 and 3, we provide a brief introduction of two Internet marketing application domains used in this study. Section 4 describes our data and data collection procedure. In §5, we present the proposed model, followed by empirical analysis, managerial implications, and economic simulation in §6. We conclude in §7.

2. User Profiling in Customer-Based Analysis

While technology enables the tracking of consumer activity online and the collection of vast amounts of consumer information, how to infer consumers' interests, and predict their behaviors from such information remains challenging. A typical consumer navigates across many websites repeatedly over time to fulfill her various needs. For example, in a shopping session, a consumer may search for a book at Google, then visit its Wikipedia page to get more information about the book, then go to both BarnesandNoble.com and Amazon.com to compare prices, and eventually purchase at the website that offers better shipping terms. At another time, the same consumer may become engaged in a technology exploration session triggered by a post about a new model of iPad on her friend's Facebook page. The interest in the product brings the consumer to Apple's website, and then to a discussion forum. The observed online activities of an individual consumer are thus a collection of visitations to many websites of different categories, which reflects a combination of her various interests and behavioral patterns.

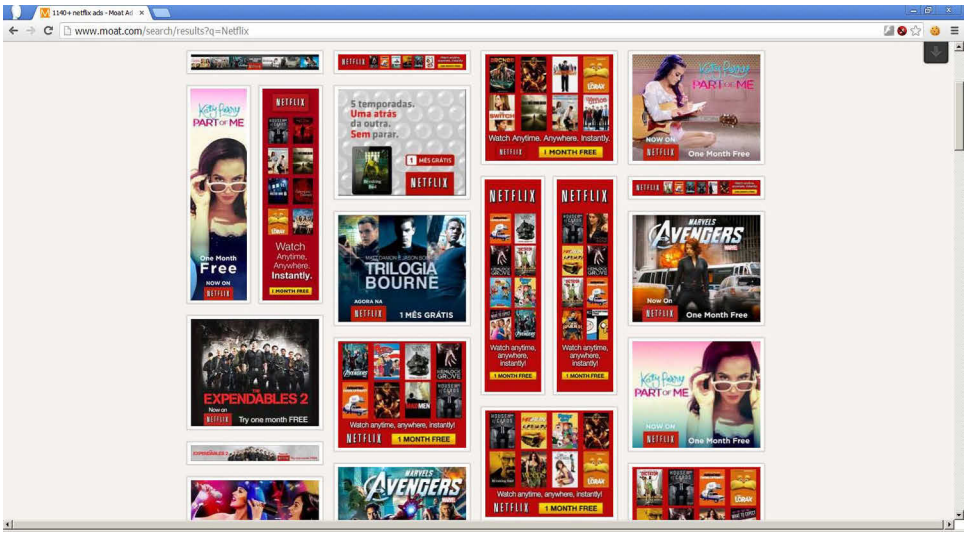
Extracting consumer interests and behavioral patterns from the multidimensional Web browsing data is an underexplored area in marketing. Early work models browsing behavior at multiple websites used the Sarmanov family of multivariate distributions (Park and Fader 2004). While an important first step, the model does not scale beyond a small number of dimensions. Danaher (2007) also draws on the Sarmanov family to develop a multivariate model with marginal negative binomial distribution, an approximation of which can be scaled up to higher dimensions. Li et al. (2015) use a multivariate normal correlation structure to explicitly connect the different categories of websites. Danaher and Smith (2011) further advance the literature by introducing the copula model, which admits flexible correlation among the multiple dimensions, and demonstrate that the model can be applied to data that cover 45 websites (dimensions).

While significant progress has been made, extant studies mostly focus on fitting the data, instead of interpreting the data and extracting behavioral patterns. Our proposed approach is better suited for this task. We model consumers' online activities as the result of a combination of different roles, each representing a distinct behavioral pattern. The model is dynamic and accounts for potential changes in individual profile composition over time. It is also scalable and works well on large data sets. Furthermore, the model yields results that are readily interpretable: a set of distinct behavioral profiles, and combinations of such profiles into individual consumers. Such interpretability is sought after by marketing researchers and practitioners alike. Meanwhile, our model admits heterogeneity, both observed and unobserved, which enables more accurate targeting when additional information of a consumer becomes available. We believe all these constitute important contributions to the literature.

3. User Profiling in Display Advertising

According to Forrester Research (Marshall 2014) display advertising (Figure 1) will account for roughly $23.7 billion in digital ad spending in 2015, second only to paid search marketing at $25.7 billion (eMarketer 2014a). The driving force behind display advertising is user information, which feeds into the ad targeting engine. There are a number of technologies enabling marketers to reach consumers with display ads. One of the most remarkable recent developments has been Real Time Bidding (RTB). In our example we focus on RTB for three reasons. First, like most other approaches to display advertising delivery, RTB makes heavy use of individual profile data. The ability to make inferences about a partially observed user profile is vital. Second, as its name suggests, profile inference must be made in real time. Third, with a variety of online businesses

Figure 1 **(Color online) Examples of Display Ads by Netflix.com**



Source. http://www.moat.com/.

using RTB technology, we can compare prediction performance across distinct sets of first-party data available to each firm.

In contrast to traditional approaches for buying display ad inventory such as static bidding or publisher bulk buying, where the advertiser pays a fixed rate for a certain number of ad impressions delivered to a target audience, with RTB advertisers can assess the potential value of each individual impression (impression scoring) and set their bid accordingly. The benefits of such impression-level optimization are obvious. Instead of relying on a predefined targeting criteria that triggers impression or submission of a preset bid (where all potential impressions that meet the criteria are valued the same), advertisers may leverage the rich information set associated with individual impression, thus increasing the effectiveness of the advertising campaign and improving return on investment (ROI). More than $11 billion of display advertising will be spent on RTB in 2015 (eMarketer 2014b).

The enabling component of RTB is the information context available to the advertiser at the time of decision making (i.e., bid submission) that is specific to each impression. A number of sources could provide this information. First, a facilitating platform (e.g., an ad exchange such as DoubleClick) allows an advertiser to set pre-targeting criteria that filters out bid requests (sent by the platform to advertisers) for impressions that are not of interest to the advertiser.[3] Hence, receiving a request for a bid is an indication that the impression meets pre-targeting criteria (e.g., an impression will be delivered through a certain publisher, the user meets a certain demographic or behavioral profile, etc.). Second, the facilitating platform may pass along some additional information about the impression (e.g., publisher website information, ad on-page position, device type, geo location, and user data, etc.). Finally, enclosed in the bid request is the platform's user ID (i.e., cookie associated with the impression) that the advertiser may match against their own cookie tables to retrieve their internal information on this user (Figure 2).

Because the facilitating platform's information is being made available to all auction participants, true competitive edge comes from the advertiser's ability to leverage their internal information on the user matched through cookie ID. As we argued above, advertisers have different amounts of data on users. At one extreme, the advertiser is unable to match platform's user ID with their internal records and, therefore, is limited to the information passed to them by the platform. At the other extreme, the advertiser may have a complete[4] user profile collected internally or purchased through third party data providers. In the latter case, the advertiser is in a much better position to score the impression and to optimally set the bid.[5]

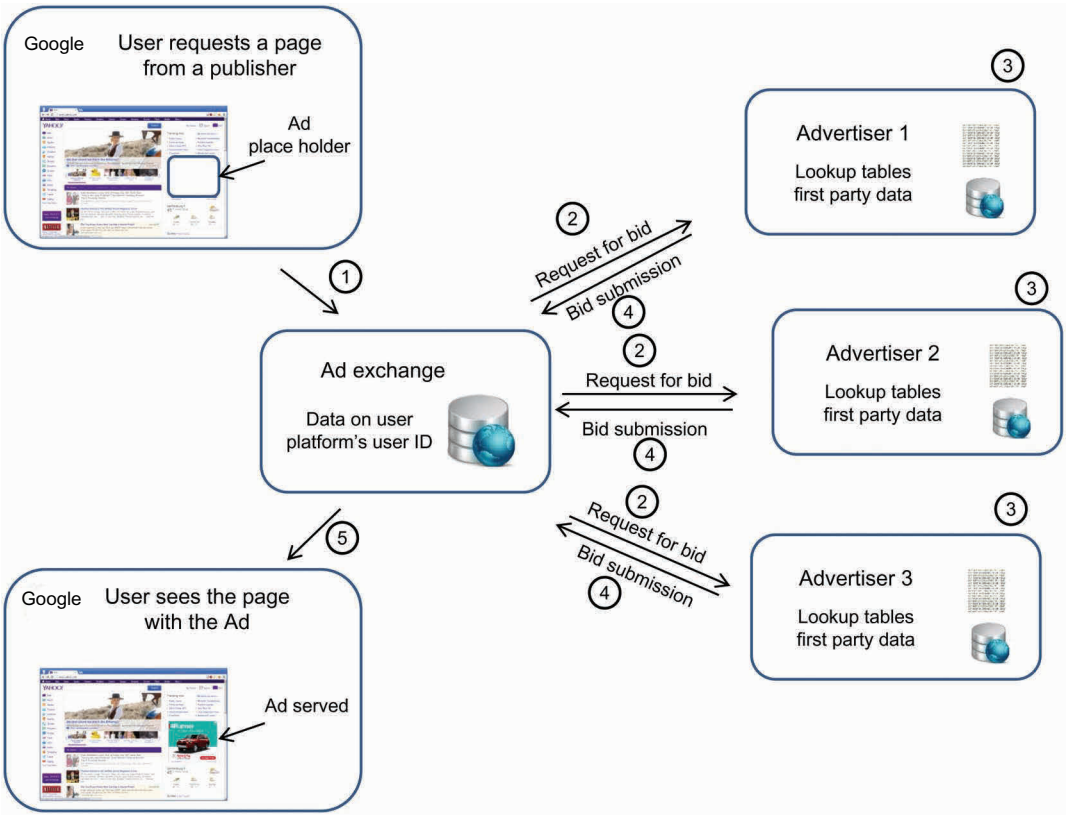
Availability of complete profile information, however, is rare. Typically, the advertiser has a fragmented view of an individual profile. For example, BarnesandNoble.com has a complete record of user's browsing

[3] In some sense pre-targeting is similar to static bidding where targeting criteria is set in advance. For more information on pre-targeting see https://developers.google.com/ad-exchange/rtb/target-guide.

[4] By the term complete we mean having unbiased information on all dimensions of user interests and preferences used to describe the user in a given targeting platform.

[5] Note that the impression scoring model is beyond the scope of this research. Instead, our model produces one of the key inputs to the scoring procedure.

Figure 2 (Color online) Simplified View of Real-Time Bidding Interactions



Notes. (1) User requests a page from the publisher's website; (2) Ad Exchange identifies user by cookie ID, pulls user's profile, selects advertisers based on their pre-targeting conditions and sends out requests for bids; (3) Advertiser receives the request for bid, matches user's ad exchange ID with its own records, and pulls user's information to score the impression; (4) Advertisers submit their bids for impression; (5) Display ad of a winning bid is served to the user along with publisher's content. Note that different kinds of online businesses may serve the role of advertiser. Demand-side platforms (DSPs) are, perhaps, the most common bidders on ad exchanges that serve as middlemen between the firm-advertiser and publishers. To streamline the presentation we omit details about the underlying multilevel interactions.

activities and book purchase history on their site.[6] However, it does not have information on other dimensions of user's interests that are not revealed through the book shopping process (e.g., interest in local services or the user's involvement in social media). Knowing these hidden dimensions of a user profile could help BarnesandNoble.com to more accurately assess the value of reaching this user through a display ad. Even for major online firms with a strong audience reach (i.e., number of unique visitors per month), profile fragmentation presents a challenge. The ability to *reach* a large share of the Internet population does not imply having comprehensive knowledge of this population. This premise may explain investors' enthusiasm for popular social media outlets (such as Facebook, Twitter or Pinterest) that are striving to achieve broad reach and consumer knowledge.

Our proposed approach can assist online advertisers in predicting the missing dimensions of profile information as well as correcting for biases that could exist in the first-party data. In the subsequent sections we show how each user profile can be reconstructed using first- and third-party data sources. We also discuss how the proposed approach can be adopted to perform in real-time making profile updates based on the data received from an ad exchange within a request for bid.

4. Data

Our data is obtained from a leading global information and measurement company that wishes to remain anonymous. The data set has detailed website browsing information on a panel of more than 45,000 households from January 2012 to December 2012. For each household in the panel, a detailed log of browsing activities at the session level is recorded. For each instance of browsing activity, information on the exact time of the webpage visit, the domain name, and the duration of

[6] User data collected internally by the advertiser is called first-party data. One of the popular behavioral targeting applications that rely primarily on first-party data is retargeting (e.g., Lambrecht and Tucker 2013). In retargeting, delivery of the display ad through a third-party publisher is triggered by the condition linked to the past user's actions on the advertiser's website (e.g., recent visit to the advertiser's website or product placement to the shopping cart).

Table 1 Category Names and Monthly Household Website Visits

Category	Mean	SD	Min	Max
Automotive	1.52	8.63	0	869
Business to business	0.66	3.89	0	309
Business/Finance	7.34	39.22	0	2,999
Career services and development	1.93	9.58	0	1,064
Corporate presence	10.16	24.83	0	1,612
Directories/Resources	9.16	21.20	0	1,725
Education	4.64	15.66	0	1,393
Entertainment	34.76	89.95	0	4,993
Family and youth	4.18	15.29	0	957
Gambling	0.82	6.72	0	520
Games	10.42	35.08	0	1,829
Government	1.79	6.04	0	386
Health	2.33	11.56	0	974
ISP	1.05	7.03	0	533
Lifestyles	7.20	28.86	0	3,859
News/Information	19.97	229.54	0	15,068
Portals	21.46	43.24	0	1,904
Promotional/Servers	34.10	73.78	0	6,452
Real estate	0.98	6.43	0	585
Regional/Local	1.23	7.81	0	1,019
Retail	16.96	36.16	0	2,474
Search/Navigation	8.71	23.16	0	3,761
Services	42.69	66.07	0	2,682
Social media	37.09	65.95	0	2,123
Sports	4.70	30.26	0	1,743
Technology	4.85	34.04	0	1,898
Telecommunications	1.41	6.77	0	405
Travel	2.19	12.29	0	1,934
XXX adult	21.66	71.75	0	3,799
Number of households: 45,300				

the visit is maintained. Because the data covers the entire Web browsing history for the households in the panel, it has more than half a million domain names (websites). These websites are organized in the data set using a two-level category classification system supplied by the collaborating marketing research firm, whereby each website visited is assigned a unique category and subcategory. There are 29 categories, and 116 subcategories. The names of all categories are listed in Table 1.

While the practice of tracking a user's online activity to build a behavioral profile is widespread, to our knowledge there is still no common standard on the taxonomy of profile data. Advertising networks typically offer advertisers a set of targeting options linked to the interests and preferences revealed by potential targets through their online activities (Figure 3). These targeting dimensions however tend to differ from one network to another. In addition, some networks support custom targeting criteria that are specific to an advertiser.

Another open issue is how the user's interests in different categories are quantified. Some ad networks offer high/low-type targeting criteria indicating interest in corresponding categories. With this dichotomized approach the target profile is represented as a binary vector. Alternatively, each dimension of the behavioral profile may be associated with a visit frequency by the user to a given class of websites. In the absence of a common standard on profile composition, we adopted the taxonomy used by the collaborating marketing research firm and represent individual profile as a vector of count data that captures the number of visits to different categories of websites. This approach to profile composition is flexible and extends well to other types of data beyond site visitation (e.g., purchases).

As our study involves profiling consumer's interest and behavior, we focus on the category composition of household website visits.[7] The monthly shares of website visits by category are presented in Figure 4.[8] The figure shows that "Entertainment," "Promotional/Servers," "Services," and "Social Media" are the most frequently visited categories, each accounting for roughly 10% of the total visits. In the next tier of categories are "News/Information," "Portals," "Retail," and "XXX Adult," each accounting for 5%–7% of total visits. The other categories that each accounts for at least 2% of total visits are "Business/Finance," "Corporate Presence," "Directory/Resources," "Games," "Lifestyles," and "Search/Navigation." Household-level descriptive statistics for monthly website visits by category are reported in Table 1. Consistent with the discussion above, entertainment, promotional servers, services, and social media are the top categories of visits. Table 1 also shows a wide dispersion of website visits across households, where all categories have maximum visits more than 10 times the population means, and large standard deviations.

The data set also has the demographics information of each household, including age and household income. Descriptive statistics are reported in Table 2. As Table 2 shows, the data set provides a balanced coverage of every strata of the population by these demographic dimensions. For age, slightly more households are between 40 and 54 years old than other ages. For income, the mode is \$50,000–\$74,999. All age and income brackets are well represented in the data set. Our proposed modeling approach accounts for such demographics information in the characterization and prediction of consumers' website visits, which we discuss in §5.

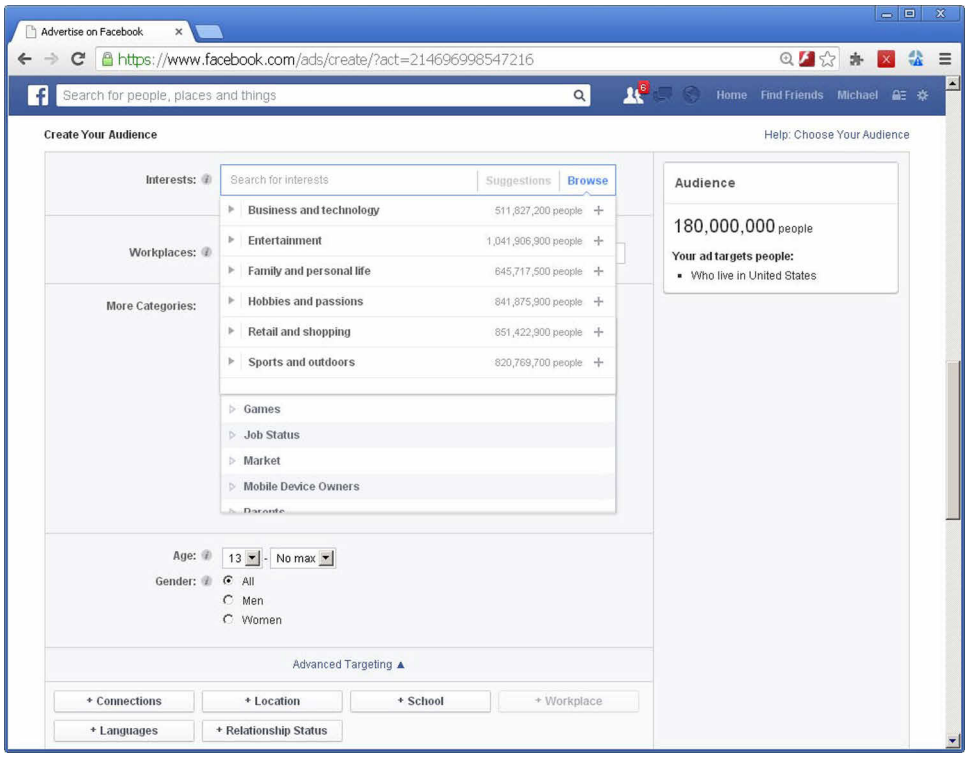
4.1. Search Engine's and Advertising Network's Perspective

Central to the issue of consumer behavioral profiling is the question of "who knows what." In the Internet

[7] For ease of exposition when discussing the data, model properties, and findings, all discussions in this paper are on the category level. Details of analyses at the subcategory level are available from the authors on request.

[8] Categories that account for more than 2% of visits are displayed in Figure 4.

Figure 3 **(Color online) Targeting Options of Facebook Ads**



ecosystem, two types of businesses, search engines and advertising networks, have emerged as holders of large amounts of information on consumer online surfing behavior and they warrant particular attention. These companies are also central to the applications of customer-base analysis and display advertising. Major search engines, such as Google and Yahoo!, function as common entry points of consumer Web visits. Consumers routinely go to a search engine, type in keywords of interest to search, and continue the Web visits by clicking on the links that are returned. Accordingly, for each search the search engine knows the link clicked by the consumer, hence the website she visits immediately after the search (i.e., the downstream site). Taken together these visit records form a search engine's knowledge of consumers' web visit behavior. The second type of online business are advertising networks, such as DoubleClick and Quantcast. Advertising network firms contract with large sets of websites to manage their display advertising inventory, i.e., spaces

Figure 4 **(Color online) Household Website Visit Histogram by Category**

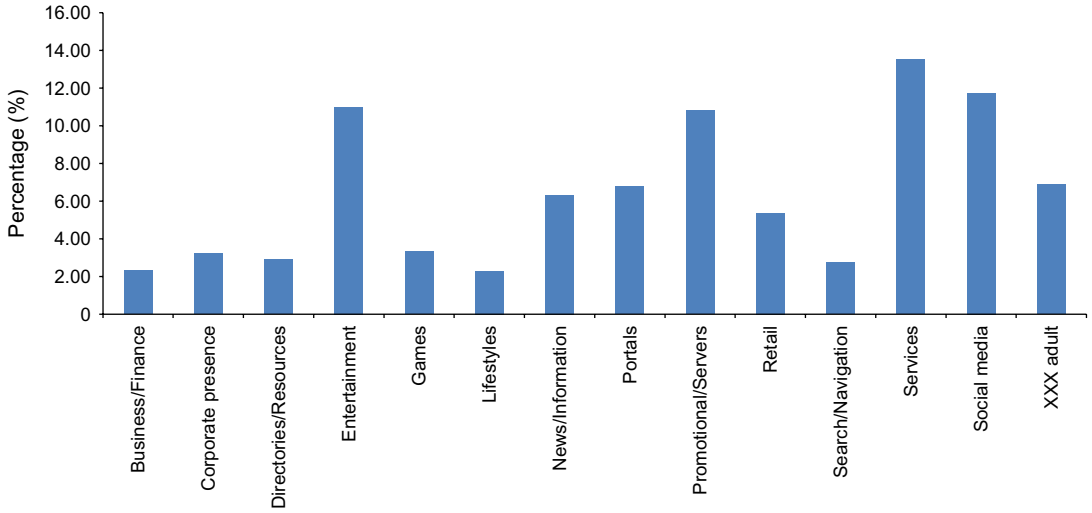


Table 2 Descriptive Statistics—Household Demographics

Age		Income	
Range	Number of households	Range	Number of households
18–20	1,815	<$15,000	7,801
21–24	2,714	$15,000–$24,999	5,401
25–29	3,569	$25,000–$34,999	5,585
30–34	4,344	$35,000–$49,999	6,004
35–39	3,881	$50,000–$74,999	9,375
40–44	5,067	$75,000–$99,999	5,391
45–49	5,674	$100,000+	5,742
50–54	5,332		
55–59	4,297		
60–64	3,097		
65+	5,509		

on the webpages where ads can be displayed. As such, an advertising network knows about consumers' web visits whenever such visits occur at the websites covered by the advertising network. Both search engine and advertising network firms, therefore, have extensive information on consumer online activities, from which they can infer consumer interests and behavior. Although they collect a large amount of information of consumer visits, their data carries certain biases. We explore these in this study.

The panel data used in our study offers a detailed picture of user level activities across a large set of online properties. Thus we can single out any of the online firms of interest and to reconstruct the corresponding firm's view on their user base (subject to sample size). It is straightforward to define information sets available to each firm as all records of individual activities that are on the firm's website. User activities on all other sites serve as the anonymized third-party data. For a search engine, the website visit immediately after a search is also observed by the search engine and included in its information set. Advertising networks present a challenge, however, as they track each user across numerous websites. Yet data on the sites participating in the network is proprietary information that networks typically do not disclose. To solve this problem, we developed a web crawler application that visits the websites found in our data set combined with 1 million websites identified by Alexa.com as top traffic sites. Our goal is to extract information about the tracking tools (including ad network tags) active on each site. To identify tracking tools we use a popular privacy software, Ghostery.[9] Given the large scale of the crawling task, we had to modify the original Ghostery code to run within our custom-built virtual browser (which does not require user interface, but which simulates a full browsing experience by downloading and executing client-side scripts) and to handle a variety of special conditions such as nonresponding or possibly malicious websites. We used cloud computing facilities provided by Amazon Web Services to run four instances of a virtual Web browser for roughly three months to collect tracking data. From this data collection effort, we reconstructed approximate coverage for all prominent multisite Internet companies.

[9] "Ghostery is a browser extension, which [...] can monitor all of the different web servers that are being called from a particular web page and matches those with a library of data collection tools (trackers)." https://www.ghostery.com/support/faq/ghostery-add-on/how-does-ghostery-work/.

Table 3 Advertising Network and Search Engine Coverage

	Domains covered	Domain coverage ratio (%)	Visit coverage ratio (%)
Advertising network			
DoubleClick	25,198	4.49	19.21
Google adsense	17,886	3.19	3.28
Facebook social plugins	15,559	2.77	4.31
Quantcast	7,006	1.25	4.29
Omniture	3,962	0.71	5.30
Twitter button	6,718	1.20	2.51
Statcounter	2,781	0.50	0.25
BlueKai	1,374	0.24	1.50
Search engine			
Google	181,501	32.33	12.18
Yahoo!	99,943	17.80	5.68
MSN	35,087	6.25	1.82
AOL	30,498	5.43	1.16
Total domains visited	561,452		

The coverage ratio of a set of large advertising networks and major search engines are reported in Table 3. Table 3 shows that large advertising networks and search engines indeed possess large amounts of consumer information. The largest advertising network according to the data set, DoubleClick, covers 4.49% of the websites in the data set, and 19.21% of consumers' web visits. Similarly, the largest search engine, Google, covers 32.33% of the websites, and 12.18% of all visits. That a single ad network knows one of every five website visits of each consumer is quite remarkable, as is the fact that a single search engine can reach one-third of all websites. Furthermore, comparing advertising networks with search engines, we see that the former tend to have higher visit coverage ratios, while the latter have higher domain coverage ratios. This is understandable, as advertising networks are embedded in websites and can track all activities. Note however that they may track only websites with significant ad inventories, and thus will not reach many smaller websites. By contrast, search engines rely on consumers going to their own websites to begin with, and can track only the first click out of the search engine but not subsequent direct visits. Because information search is diverse, however, they may reach a larger set of websites, big and small. Thus, although search engines and

advertising networks cover a large number of websites, their perspectives on consumer visits are somewhat different. Finally, neither advertising networks nor search engines seem to have a proportionally accurate picture of consumers' website visits. This is illustrated in Figures 5(a) and 5(b), which show the category histogram for major search engines and advertising networks, respectively. Figure 5(a) shows that all four search engines have category histograms different than the actual consumer visit profiles. For example, Google over-counts the share of Directory/Resources while it under-counts that of Services. As another example, MSN over-counts Corporate Presence while it under-counts Entertainment. The histograms also differ significantly across the search engines, suggesting that each is holding a somewhat unique information set. Similarly, Figure 5(b) shows that major advertising networks also have disproportionate pictures of consumer website visits. Such biases differ from network to network. This bias appears to be somewhat more severe than search engines. For example, the share of Entertainment and Portals categories from DoubleClicks is about three times their share in the actual consumer profile. This exploratory analysis highlights the limitations of the first-party data and calls for improvement in consumer behavioral profiling.

5. Model

In this section, we present the model that we use to characterize consumers' website visit activities. Our model extends from the CTM model (Blei and Lafferty 2007) by incorporating visitation intensity, heterogeneity, and dynamics. CTM is among the latest developments in the family of Topic Models that have received widespread attention in fields such as machine learning and natural language processing. Topic models were originally used to identify the mixture of topics present in a large number of documents. Just as a document can be considered a combination of multiple topics, the key insight of topic models, a consumer's website visit activities can be considered the combination of multiple roles or objectives. For example, the consumer may play a social role, where she visits sites such as Facebook.com or Twitter.com; she may play a shopper role at another time, where she visits places such as Ebay.com or Amazon.com; she may also play an information seeker role at yet another time, where she visits CNN.com and blogs, etc. Therefore, topic models are a good conceptual fit to our task of user profiling using website visit data.

Topic Models originated from probabilistic latent semantic indexing (Papadimitriou et al. 1998, Hofmann 1999), and have evolved over the years. The most commonly used model in this family is the Latent Dirichlet Allocation (LDA) (Blei et al. 2003, Griffiths and Steyvers 2004). A generative statistical model, LDA models the generation of documents that are composed of multiple topics or sources in two steps, from document to topic, and then from topic to word. Each step is a multinomial random draw with Dirichlet prior, hence the name. LDA has received widespread acceptance due to its power to extract meaningful topics using a parsimonious model, and is the de facto standard of Topic Models.

Although popular, the LDA model suffers from a major limitation, i.e., its inability to model correlations in topic composition (Blei and Lafferty 2007). Because of the Dirichlet prior, the proportion of one topic cannot be explicitly correlated with that of another. In many real world applications, such a correlation is expected. For example, for roles (i.e., topics) underlying website visits, we would expect the proportion of the role business taken by an individual to be more positively correlated with the role career development than with the role entertainer.[10] The CTM model was developed to address this limitation. It uses a multivariate normal prior for document-topic composition to explicitly model correlations.

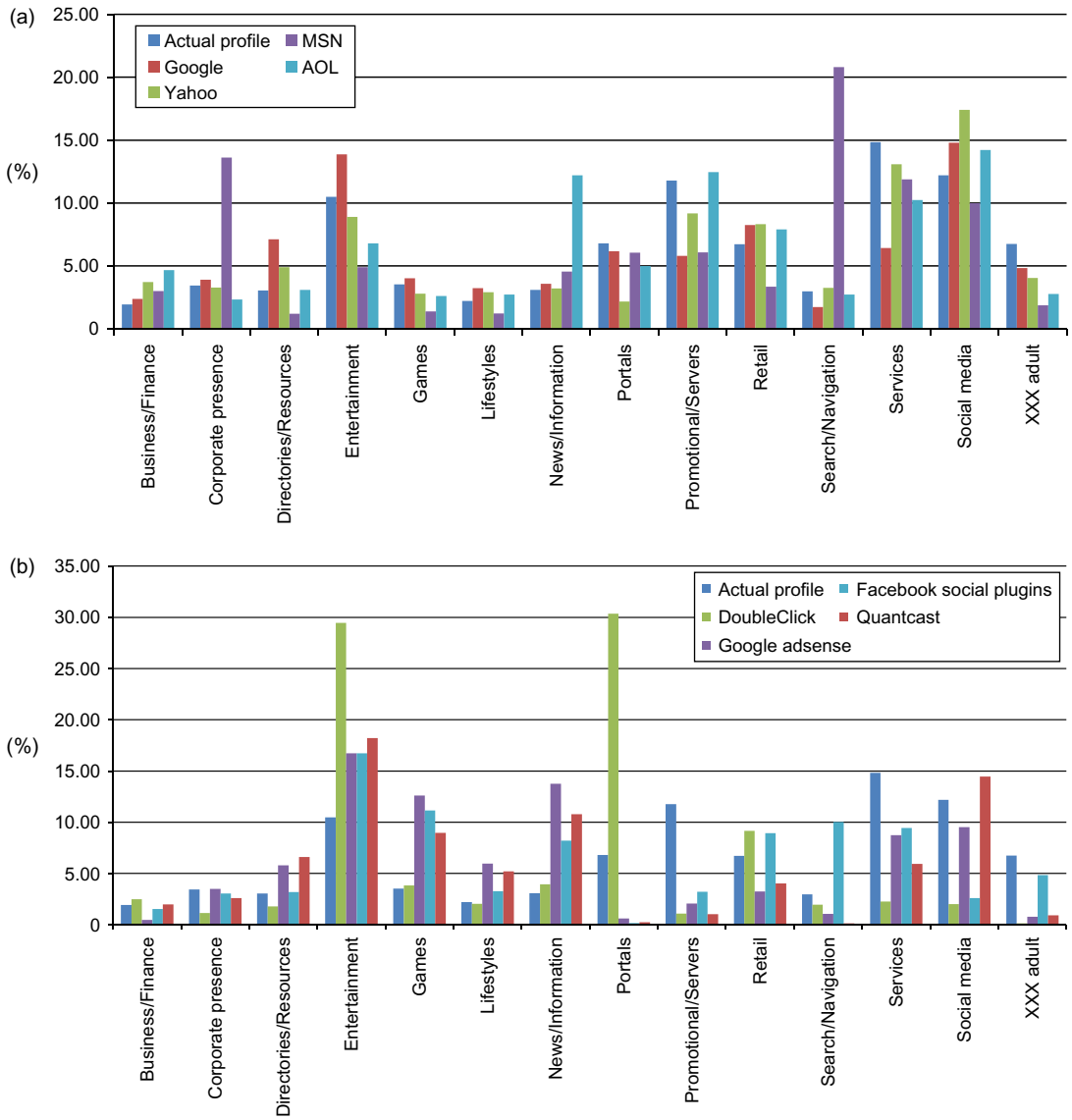
Our model is developed along the lines of CTM. A consumer's website visitation profile can be considered as a document. A consumer's profile is a combination of different roles. For example, the consumer may use the Internet for entertainment on some occasions and for shopping at other times. These roles correspond to topics in Topic Models. While our model is grounded in the standard CTM, we make three important extensions. First, in our model we account for website visitation intensity, i.e., the activeness of the user during Web surfing, and its correlation with roles. The intensity corresponds to document length in Topic Models. (Document length is commonly taken as exogenously given.) Second, our model includes explanatory variables such as consumer demographics. This better admits heterogeneity and gives a richer description of consumer profiles. Finally, we model the evolution of roles over time to account for changes in consumers' website visitation tendencies.

We now formally set up our model. There are I consumers, each indexed by i, $i = 1, \ldots, I$. There are T time periods, each indexed by t, $t = 1, \ldots, T$. For each consumer at each time period, we observe a vector of category-specific website visit counts. The website visit profile of consumer i at time t is denoted as

$$V_{it} = (V_{it1}, \ldots, V_{itC}). \tag{1}$$

[10] Part of this correlation can be seen from underlying demographics. For example, a middle-aged online user may be more inclined to career- and business-related content, while a younger person is interested in social networking and entertainment. Yet because it may also reflect latent user traits, it is necessary to explicitly model such correlation.

Figure 5 **(Color online) Visit Profile Histogram—Search Engine's and Ad Networks Perspective**



In the vector, V_{itc} is the number of times a consumer visits websites that belong to category c in the time period t. The total number of categories is denoted as C.

Following the conceptual framework of Topic Models, we assume that each consumer visit is a *two-step process*. First, the consumer decides on the *role* for the website visit. Next, according to the role selected in the first step, the consumer decides on the *website* to visit. For example, a consumer may decide to do some online shopping and then visit Amazon.com. Another consumer may decide to connect with friends, and then visit Facebook and Twitter. A consumer is expected to have multiple needs, including shopping, social, education, etc. The overall consumer website visit profile reflects a combination of the different roles the consumer plays to satisfy those needs. Different consumers would have distinct emphases on individual roles. A college student, for example, may spend more time playing education and social roles than a retired person. Formally, there are R different roles. The role-composition of consumer i in time period t is denoted as

$$\mathrm{P}_{it} = (p_{it1}, \ldots, p_{itR}). \tag{2}$$

In the vector, p_{itr} is the percentage of times when the consumer plays role r, i.e., the probability of role r, in time period t.

The probability of a consumer visiting a website of certain category depends on the role the consumer is playing at that time. Someone who is shopping online may visit Amazon.com and Ebay.com, while someone who is studying may visit a university website. In our model, each role is represented as a distribution over

different website categories. The distribution of role r over the website categories is denoted as

$$\Phi_r = (\phi_{r1}, \ldots, \phi_{rC}). \tag{3}$$

In the vector, ϕ_{rc} is the probability that a consumer taking role r will visit a website that belongs to category c.

Furthermore, the total number of visits of consumer i at time t is denoted as N_{it}. We assume that the total number of visits, which indicates the consumer's Internet use intensity, is drawn from a Poisson distribution

$$N_{it} \sim \text{Poisson}(\lambda_{it}). \tag{4}$$

There are R values in P_{it}, of which $R-1$ are free parameters (the probabilities sum to 1). We perform a logit-transformation so that P_{it} is generated from parameter $\Theta_{it} = (\theta_{it1}, \ldots, \theta_{itR})$ where we normalize $\theta_{itR} = 0$ as follows:

$$\begin{cases} p_{itr} = \dfrac{\exp(\theta_{itr})}{1 + \sum_{r'=1,\ldots,R-1} \exp(\theta_{itr'})}, & r = 1, \ldots, R-1 \\ p_{itR} = \dfrac{1}{1 + \sum_{r'=1,\ldots,R-1} \exp(\theta_{itr'})} \end{cases}. \tag{5}$$

Re-parameterizing in this way allows us to introduce observed heterogeneity, unobserved heterogeneity, and dynamics of the consumer's role composition. In this set-up, θ_{itr} represents the probability of consumer i playing role r at time t, relative to role R (since θ_{itR} is normalized to 0). As an extension to the original CTM model, we model θ_{itr} as

$$\theta_{itr} = \theta_{ir} + X'_{it}\vec{\rho}_r + \delta_{itr}, \quad r = 1, \ldots, R-1. \tag{6a}$$

In Equation (6a), θ_{ir} is consumer i's baseline propensity for role r. A positive value of θ_{ir} indicates that the rth role accounts for a bigger portion of website visitation than the last role, role R. The variable X_{it} is a vector of observed characteristics that can be consumer-specific, time-specific, or both. Firms with large amounts of data on such characteristics can thus leverage such information to improve user profiling accuracy. In our data set, X_{it} includes consumer demographic variables such as age and household income, and time varying covariates that capture public interest in popular search categories based on Google Trends data (http://www.google.com/trends/), and $\vec{\rho}_r$ are the corresponding coefficients. δ_{itr} is an individual- and time-specific factor that captures the evolution of the consumer's role propensities over time.

Similarly, we model the use intensity parameter as

$$\lambda_{it} = \exp(\lambda_i + X'_{it}\vec{\rho}_\lambda + \delta_{it\lambda}). \tag{6b}$$

Similar to the role composition set-up, λ_i is the baseline Web visitation intensity of consumer i. The term $X'_{it}\vec{\rho}_\lambda$ captures how the visitation intensity varies according to observed consumer- or time-specific covariates. $\delta_{it\lambda}$ is an individual- and time-specific factor that captures the evolution of the consumer's visitation intensities over time.[11]

Consumers' role composition and use intensity are expected to evolve over time. As an extension to the standard CTM model, we capture the dynamics at population and individual consumer levels, using an approach similar to Kalman filtering. Specifically, we model the time-specific factors for role composition and visitation intensity as

$$\begin{cases} \delta_{itr} = \phi_r \delta_{i,\,t-1,\,r} + \xi_{tr} + \epsilon_{itr}, & r = 1, \ldots, R-1 \\ \delta_{it\lambda} = \phi_\lambda \delta_{i,\,t-1,\,\lambda} + \xi_{t\lambda} + \epsilon_{it\lambda} \end{cases}. \tag{7}$$

In Equation (7), ξ_{tr} and $\xi_{t\lambda}$ are the population level time-specific factors that capture the overall change of role composition and visitation intensity over time. The terms $\phi_r \delta_{i,\,t-1,\,r}$ and $\phi_\lambda \delta_{i,\,t-1,\,\lambda}$ are the autoregressive terms that capture potential persistence at the individual consumer level, where $\phi_r \in (-1, 1)$ and $\phi_\lambda \in (-1, 1)$ to ensure the stationarity of the process. $\epsilon_{itr} \sim N(0, \sigma^2_{\epsilon r})$ and $\epsilon_{it\lambda} \sim N(0, \sigma^2_{\epsilon\lambda})$ are i.i.d. random disturbances.

Compared to the standard CTM model, our model further admits observed and unobserved heterogeneity. As discussed above, the observed heterogeneity is introduced through the term $X'_{it}\vec{\rho}_r$ and $X'_{it}\vec{\rho}_\lambda$. Admitting observed heterogeneity this way allows us to analyze how observed consumer characteristics and other observed characteristics determine role composition. For example, if age is observed and we expect a younger consumer to spend more time playing a social role, then the coefficient for age for the social role should be positive. Furthermore, we introduce unobserved heterogeneity by treating the individual consumer-specific baseline role composition parameters and the use intensity parameter of consumer i as drawn from a population-level multivariate normal distribution. Inclusion of the use intensity parameter is another departure from and extension to the original CTM, and helps us understand how different roles are related to the amount of Web activities

$$\begin{pmatrix} \theta_{i1} \\ \cdots \\ \theta_{iR-1} \\ \lambda_i \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\theta}_1 \\ \cdots \\ \bar{\theta}_{R-1} \\ \bar{\theta}_\lambda \end{pmatrix}, \Sigma \right). \tag{8}$$

[11] Including parameters capturing observed and unobserved heterogeneity across consumers and variation over time in Equation (6b) also helps to account for overdispersion in the data. The estimation result, discussed in detail in §6, shows that the predicted visitation counts based on our parameter estimates have a mean of 261.2 and a variance of 190,628, very close to the sample statistics of 261.3 and 190,882, respectively.

In Equation (8), Σ encodes the variance of the distribution of each role across consumers, and the correlations among roles and between roles and the website use intensity.

As discussed earlier, each consumer visit is generated from a two-step process. For each visit v, $v = 1, \ldots, N_{it}$ (N_{it} is drawn as in Equation (4)), she first decides on a role

$$r_{itv} \sim \text{Multinomial}(\mathrm{P}_{it}). \tag{9}$$

Then, based on the chosen role, she decides on the category of the website to visit

$$c_{itv} \sim \text{Multinomial}(\Phi_{r_{itv}}). \tag{10}$$

The overall visit profile is then summarized as

$$V_{it} = (V_{it1}, \ldots, V_{itC}), \tag{11}$$

where

$$V_{itc} = \sum_{v=1}^{N_{it}} I\{c_{itv} = c\}. \tag{12}$$

In Equation (12), $I\{\cdot\}$ is the indicator function.

Finally, to complete the model, we assume that the role-category mapping is drawn from a Dirichlet distribution, as in the original LDA model

$$\Phi_r \sim \text{Dir}(\vec{\alpha}). \tag{13}$$

The model can be estimated using a hierarchical Bayesian approach with data augmentation. The algorithm and estimation details are provided in Technical Appendix 1 (available as supplemental material at http://dx.doi.org/10.1287/mksc.2015.0956). Our model and estimation approach are particularly suitable for a large-scale data set (i.e., big data) due to its scalability. Efficient estimation of topic models has been a focus of research. Existing studies have shown that models such as LDA can be efficiently estimated in a parallel manner. Although our model is more complex, the architecture of our model is similar to that of standard topic models from the perspective of estimation. In fact, our proposed model falls into the general framework of Latent variable models discussed in Ahmed et al. (2012), which demonstrates how this category of models can be efficiently estimated in parallel using, for example, Hadoop. To further demonstrate the scalability, we also developed a detailed parallel estimation algorithm, and tested its implementation in a multiprocessor environment. This shows that the estimation increases almost linearly when the number of CPUs increases. Processing time per profile record is stable as the size of the database grows. A more detailed discussion on scalability is provided in Technical Appendix 2.

A key issue of interest in our study is the prediction of a consumer's overall visit profile using partial information. Using the estimates of consumers' role composition and the category composition of each role, this prediction can be conducted through standard Bayesian update. Assume that we know a subset of consumer i's website visits, denoted as $\tilde{V}_i$. The prediction task involves finding the posterior of $\theta_i = (\theta_{i1}, \ldots, \theta_{iR-1}, \lambda_i)^t$ given the data $\tilde{V}_i$. This posterior can be generated in the same way as the model is estimated, traversing the parameter space of θ_i using the Metropolis–Hastings algorithm. Using this approach, we can generate a prediction of a consumer's overall visit profile from any subset of data about the consumer. We can then evaluate the information content of a subset of data based on how well it predicts the overall profile.

6. Results

In this section we show how our approach can be used in customer-based analysis and display advertising. We use the former application to highlight the interpretability of the results produced by our method. The latter application demonstrates the profile recovery performance based on the limited data available to different online firms under a variety of business scenarios. We also discuss how the proposed procedure can be adopted to real-time environments. We conclude the section with an economic simulation that illustrates potential gains the proposed model may offer to a firm if used in individual-level targeting of display ads.

6.1. Method Application to User Profiling in Customer-Base Analysis

Our data set consists of 45,300 households and covers 12 months. We use monthly level data for analysis, i.e., each time period in the model corresponds to one month.[12] We used the first 11 months for estimation, and the remaining month as a holdout sample. We took 100,000 Markov Chain Monte Carlo (MCMC) draws, and used the latter 50,000 draws for inference.

6.1.1. Model Fit. We estimated the model with a different number of roles, ranging from 2 to 12. The analyst can determine the number of roles to strike a balance between parsimony and precision. The smaller the number of roles, the more parsimonious the model is and the more dimensionality reduction it achieves; the larger the number of roles the more precise the depiction of user behaviors. We assess model fit using log-marginal density (LMD), in-sample mean absolute percentage error (MAPE), and out-of-sample MAPE.[13] The model fit statistics are reported in Table 4.

[12] We also estimated the model using data aggregated at weekly and biweekly levels, and recovered roles and role compositions similar to those recovered using monthly level data.

[13] MAPE is calculated as the sum of the wrong predictions of category visit count (i.e., the difference between predicted number of visits to a category and the actual number of visits to that category) over all categories, divided by the total number of visits, averaged across all users.

Table 4 Model Fit by Number of Roles

Number of roles	LMD[a]	MAPE—In sample (%)	MAPE—Out of sample (%)
2	−7,790,382	70.66	71.95
3	−7,551,771	61.86	65.37
4	−7,418,350	57.23	62.44
5	−7,333,782	53.00	60.22
6	−7,206,092	49.12	57.74
7	−7,132,879	45.87	55.50
8	−7,075,614	42.94	53.94
9	−7,012,142	38.38	51.43
10	−6,955,623	35.31	49.48
11	−6,901,201	30.27	47.47
12	−6,863,833	30.23	47.14

[a]Model selection performed using a randomly selected sample of 1,000 households.

As shown in Table 4, as the number of roles increases, the LMD increases, and in-sample and out-of-sample MAPE decrease.[14] Using more roles thus leads to a more accurate description of the data, as expected. The improvement in sample fit from one more role is significant when the number of roles is small. After allowing for seven or eight roles, though, the improvement brought about by an additional role decreases. In the subsequent discussion, we interpret the results from using seven roles, which serve as a balance between precision and parsimony. Depending on whether clear and intuitive interpretation is more important than accurately fitting the data, firms using our model can use discretion to choose fewer or more roles.

We have also compared the performance of our model with more traditional customer segmentation models, such as variants of cluster analysis methods. Our model offers significant improvements over these methods. This is discussed in more detail in Technical Appendix 3.

6.1.2. Behavioral Profile—Roles. The parameter estimates for role-to-category mapping (i.e., Φ_r in the model) are reported in Table 5. In this table, only posterior means are reported for ease of exposition (as the posterior distributions of these role probability parameters are constrained to stay between 0 and 1 there is no need to verify whether the credible intervals include zero).

The table highlights the salient roles consumers play when they surf the Web. In the first role, the consumer predominantly visits websites in the entertainment, services, and games category. Accordingly, we give this role an alias "entertainer." In the second role, the consumer primarily visits search/navigation and promotional servers websites, hence the alias "information seeker." In the third role, the consumer visits websites in the services, retail, directories/resources, and corporate presence categories. Education and lifestyle websites also account for a noticeable portion of this role. This depicts an "online shopper." In the fourth role, the consumer spends more than half the time on social media, and other times on services and entertainment websites. We call this the "social networker." Similarly, the remaining roles depict "family person" (news/information, sports, entertainment), "career person" (portals/business finance/corporate presence), and "adult entertainer" (XXX adult/promotional servers), respectively.

The analysis and characterization of these roles demonstrate the benefit of the model developed in this study. From count data on multiple categories of websites, salient behavioral patterns can be uncovered. In most cases, roles are characterized by a combination of multiple categories of websites, which single-dimension models would be unable to uncover. These salient, easy-to-interpret patterns facilitate the understanding of consumers' web visit activities and, in turn, their underlying interests and intentions. These factors are of keen interest to researchers and practitioners.

6.1.3. Composition of Roles. A consumer plays different roles at different times and occasions. In addition to uncovering distinct consumer behavior patterns conveniently encoded in the roles, our model also enables rich characterization of the composition of such roles. The parameter estimates for population level role-composition parameters ($\bar{\theta}$ and Σ in the model) are reported in Tables 6 and 7, for mean and variance, respectively.

The composition of roles can be more intuitively interpreted using the role percentages and correlation matrix to which they correspond, which are reported in Tables 8 and 9, respectively. (The parameters estimated in the model correspond to the roles relative to the last role, which is normalized for identification, and can be interpreted on a relative but not absolute basis.) As Table 8 shows, the third role, online shopper, is the most frequently played, accounting for 33.85% of the Web visit activities. The next most often played roles are social networker, information seeker, and entertainer. The correlation matrix in Table 9 further shows how these roles are related. Because roles are, by default, substitutes, most of the correlations are negative. Still, their magnitudes are informative and generate the following observations: (1) The social networker role is strongly negatively correlated with all other roles, suggesting that consumers who play the social networker role are likely different from those who play the other roles; (2) The roles information seeker, family person, and career person are only slightly negatively correlated, suggesting that the same people may switch among these roles at different times; (3) The

[14] As a robustness check, we also assessed the out-of-sample fit using another set of consumers instead of the last month as the holdout sample. The results, consistent with those reported in Table 4, are available from the authors on request.

Table 5 Parameter Estimates—Role to Category Mapping

Category	Role 1 (Entertainer)	Role 2 (Information Seeker)	Role 3 (Online Shopper)	Role 4 (Social Networker)	Role 5 (Family Person)	Role 6 (Career Person)	Role 7 (Entertainer—Adult)
Automotive	0.05	0.29	1.85	0.00	0.07	0.28	0.23
Business to business	0.06	0.31	0.63	0.00	0.00	0.23	0.05
Business/Finance	0.00	1.01	3.09	0.00	7.52	7.18	0.00
Career services and development	0.00	0.49	2.13	0.00	0.07	1.06	0.00
Corporate presence	1.80	2.08	8.92	0.53	1.68	4.91	0.23
Directories/Resources	1.88	0.00	8.16	2.30	0.34	3.12	0.64
Education	0.30	0.00	6.23	0.57	0.06	0.00	0.00
Entertainment	34.22	0.00	1.22	16.38	10.81	0.00	3.31
Family and youth	6.30	0.45	0.69	0.00	0.05	0.33	0.00
Gambling	0.00	1.83	0.00	0.00	0.00	0.58	0.00
Games	18.47	0.77	0.00	0.00	0.00	0.00	0.00
Government	0.00	0.15	1.93	0.02	0.06	1.38	0.00
Health	0.09	1.21	1.93	0.00	0.76	1.07	0.00
ISP	0.01	0.32	0.14	0.00	0.00	2.65	0.00
Lifestyles	2.18	0.80	4.63	1.40	1.89	2.99	0.60
News/Information	0.18	0.00	0.55	0.33	53.48	0.74	0.01
Portals	0.95	0.00	0.36	1.01	2.28	61.01	0.01
Promotional/Servers	7.38	64.15	1.99	3.33	0.00	0.56	14.05
Real estate	0.00	0.00	1.21	0.00	0.00	0.62	0.00
Regional/Local	0.05	0.19	0.84	0.06	1.15	0.45	0.00
Retail	2.47	0.58	18.75	1.63	0.02	4.75	0.79
Search/Navigation	0.17	21.95	0.37	0.00	0.29	2.15	0.00
Services	20.27	0.00	28.04	17.46	0.14	0.00	6.03
Social media	1.76	2.51	0.74	54.20	1.51	0.00	0.01
Sports	0.01	0.00	0.45	0.33	10.42	0.72	0.66
Technology	1.34	0.40	1.59	0.39	7.29	0.09	0.17
Telecommunications	0.03	0.33	0.96	0.08	0.00	1.84	0.00
Travel	0.00	0.17	2.61	0.00	0.10	1.27	0.00
XXX adult	0.00	0.00	0.00	0.00	0.00	0.00	73.22

Note. All numbers reported are posterior means in percentage.

role of career person is strongly negatively correlated with social networker and entertainer, suggesting an underlying theme of business instead of leisure. The correlation of roles with visit intensity is also revealing: The entertainer, social networker, and family person roles are positively correlated with visit intensity, while the information seeker, shopper, and career person roles are negatively correlated with that intensity. This suggests that consumers tend to have more website visits when they are in leisure mode, but fewer website visits for work or home duties.

6.1.4. Effect of Demographics. The effect of consumer demographics on role composition and visit intensity is reported in Table 10, which shows that consumer demographics can further inform as to their web visit activities. A few interesting observations can be made. First, the coefficient for age is negative for the entertainer role and social networker role. This is consistent with the profile of young adults spending time on entertainment and social media websites. Second, higher income consumers are more likely to play the online shopper role. This is also reasonable as higher disposable income enables spending. Third, the older a consumer is and the higher the consumer's income, the more likely she will play the family person role. Finally, consumers with higher income tend to have lower visitation intensity.

In summary, our model estimates provide a rich characterization of consumers' web surfing activities. Salient behavioral patterns are extracted and encoded in different roles. The unobserved and the observed heterogeneity further describe in detail the relationship among roles in their composition, and the relationship

Table 6 Parameter Estimates—Role Composition—Mean $(\bar{\theta}_1, \ldots, \bar{\theta}_{T-1}, \ldots, \bar{\lambda})$

Parameter	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
Role composition ($\bar{\theta}_7 = 0$)				
$\bar{\theta}_1$	1.644	0.015	1.616	1.674
$\bar{\theta}_2$	1.843	0.022	1.802	1.884
$\bar{\theta}_3$	2.606	0.020	2.568	2.643
$\bar{\theta}_4$	2.037	0.029	1.982	2.088
$\bar{\theta}_5$	0.338	0.027	0.289	0.388
$\bar{\theta}_6$	1.588	0.021	1.548	1.629
Visit intensity				
$\bar{\lambda}$	5.154	0.007	5.140	5.167

Table 7 Parameter Estimates—Role Composition—Covariance Matrix

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	λ
θ_1	8.041 [7.824, 8.236]						
θ_2	5.935 [5.794, 6.087]	8.684 [8.515, 8.879]					
θ_3	5.694 [5.556, 5.832]	7.341 [7.176, 7.524]	9.240 [9.046, 9.452]				
θ_4	5.269 [5.146, 5.394]	5.862 [5.724, 6.012]	6.239 [6.097, 6.398]	7.385 [7.249, 7.519]			
θ_5	5.779 [5.635, 5.928]	7.113 [6.928, 7.303]	7.702 [7.498, 7.907]	6.038 [5.871, 6.216]	10.160 [9.747, 10.532]		
θ_6	5.597 [5.444, 5.742]	7.672 [7.498, 7.847]	7.741 [7.570, 7.915]	6.170 [6.013, 6.327]	7.725 [7.522, 7.919]	10.498 [10.264, 10.714]	
λ	−0.392 [−0.436, −0.349]	−0.969 [−1.014, −0.923]	−1.008 [−1.055, −0.963]	−0.348 [−0.389, −0.307]	−0.320 [−0.370, −0.272]	−1.084 [−1.132, −1.037]	2.095 [2.066, 2.123]

Note. Numbers in brackets are 95% credible intervals.

Table 8 Average Role Probabilities

	Role1	Role2	Role3	Role4	Role5	Role6	Role7
Probability of roles (%)	12.94	15.79	33.85	19.18	3.51	12.23	2.50

Note. Probabilities of roles calculated from posterior means of the role probability parameters.

between consumer characteristics and their role compositions. Insights derived from the model estimates enhance our understanding of online activities, and affirm the power of the model developed in our study.[15]

6.1.5. Evolution of Roles. Consumer's role compositions and visitation intensities may evolve over time. The estimates of time-specific fixed effects are reported in Table 11. The parameter estimates show that at the population level role composition does evolve from month to month. The evolution is rather gradual, though, as most estimates are close to zero. We make a few observations: First, visitation intensity gradually decreases through the months, which is consistent with the descriptive statistics. Second, the weights of online shopper and family person roles increase somewhat towards the end of the year, potentially reflecting the approach of the holiday season. Third, the weights of information seeker and social networker increase during the summer months, potentially reflecting increased social activities at that time. The coefficient estimates of the individual level auto-regressive parameters are reported in Table 12. These coefficients show that at the individual consumer level, role composition and visitation intensity are persistent over time: All autoregressive coefficients are higher than 0.8. Meanwhile, the role composition and visitation intensity change meaningfully month to month at the individual consumer level, as the variances of the random errors are fairly large for all roles and for visitation intensity, especially compared with the population level parameter estimates. This suggests that, although at the population level any change in role composition is gradual, at the individual level such composition can be fairly volatile.[16]

6.2. Method Application to Profile Predictions in Display Advertising

In this section we focus on individual profile predictions assuming the positions of different online firms that operate on diverse information sets. We focus on two types of companies, i.e., search engines and advertising networks. In our hypothetical scenario we assume that these firms are bidders in a RTB auction run by an ad exchange. In this auction all of these firms serve as advertisers who need to make a prediction about the profile of an upcoming impression opportunity using three pieces of information, i.e., first-party data unique to each advertiser, anonymized third-party data available to all advertisers, and ad exchange user data passed along with a bid request. Figure 6 depicts our approach to inference procedure. The first-party data available to each of these firms is discussed in §4; the prediction method is discussed in §5. We evaluate the profile prediction performance based on how well predictions are made on the number of visits on each category of websites by each consumer.

[15] In addition to the consumer demographics variables, we included three time-specific predictors, which are the first three principal components of the Google trends of top categories. Their coefficients are reported in Technical Appendix 4.

[16] We also compared the model fit measures of our dynamic model with that of a static model, i.e., one wherein the role composition does not change over time. The LMD, in-sample MAPE, and out-of-sample MAPE of the static model are −7,968,972, 51.13%, and 62.15%, respectively. These results are significantly inferior to the dynamic specification.

Table 9 Role Composition—Correlation Matrix

	Role1	Role2	Role3	Role4	Role5	Role6	Role7	Visit intensity
Role1	1.000	−0.158	−0.355	−0.161	−0.117	−0.262	−0.101	0.113
Role2	−0.158	1.000	−0.127	−0.229	−0.087	−0.043	−0.169	−0.136
Role3	−0.355	−0.127	1.000	−0.272	−0.065	−0.158	−0.272	−0.191
Role4	−0.161	−0.229	−0.272	1.000	−0.132	−0.188	−0.133	0.089
Role5	−0.117	−0.087	−0.065	−0.132	1.000	−0.039	−0.086	0.138
Role6	−0.262	−0.043	−0.158	−0.188	−0.039	1.000	−0.169	−0.089
Role7	−0.101	−0.169	−0.272	−0.133	−0.086	−0.169	1.000	0.139
Visit intensity	0.113	−0.136	−0.191	0.089	0.138	−0.089	0.139	1.000

Table 10 Parameter Estimates—Effect of Demographics on Role Composition

	Age	Income
Role1	−0.052[a] [−0.055, −0.049]	−0.005[a] [−0.007, −0.003]
Role2	0.046[a] [0.043, 0.049]	−0.122[a] [−0.124, −0.119]
Role3	0.001 [−0.003, 0.003]	0.064[a] [0.061, 0.066]
Role4	−0.063[a] [−0.065, −0.059]	−0.053[a] [−0.055, −0.051]
Role5	0.051[a] [0.046, 0.056]	0.088[a] [0.086, 0.090]
Role6	0.170[a] [0.166, 0.177]	−0.033[a] [−0.035, −0.030]
Visit intensity	0.003 [0.002, 0.005]	−0.021[a] [−0.022,−0.021]

Note. Numbers in brackets are 95% credible intervals.
[a]95% credible interval does not include zero.

As discussed in §4, although advertising networks and search engines have a considerable amount of user information, their perspectives are biased: Certain categories are overrepresented while others are underrepresented. As shown in Table 13, the original biases, measured as the absolute difference between the category level histogram of the data available to each advertising network or search engine and that of the actual profiles, range from 0.303 for Yahoo! to as high as 0.965 for DoubleClick. For predictions made using our proposed model, by contrast, the biases are reduced significantly for all of the advertising networks and search engines. Most companies achieve an improvement of more than 40% (with the exception of AOL and Yahoo! whose original bias is low). This bias reduction comes from both layers of our model, i.e., mapping consumers into a combination of roles, and using roles to connect multiple website categories. Both are generated from third-party anonymized data and serve as common priors in this prediction. Note that this reduction does not necessarily benefit those

Table 11 Parameter Estimates—Time Fixed Effects

Month	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_λ
January							
February	−0.041	−0.017	0.055	−0.066	−0.184	−0.080	−0.232
March	−0.026	−0.223	−0.104	−0.367	−0.007	−0.097	−0.169
April	−0.084	0.040	−0.008	0.169	−0.240	−0.011	−0.191
May	−0.008	−0.087	−0.044	−0.138	−0.057	−0.107	−0.216
June	−0.051	0.071	−0.066	0.173	−0.300	0.063	−0.222
July	−0.054	0.050	0.029	0.187	−0.128	0.071	−0.161
August	0.038	−0.135	0.072	−0.270	0.156	0.049	−0.137
September	0.001	0.073	0.081	−0.135	0.212	−0.146	−0.245
October	0.246	0.054	0.162	−0.161	0.299	−0.008	−0.156
November	−0.155	0.045	0.008	0.279	−0.163	0.002	−0.307

Notes. All numbers reported are posterior means. January is normalized to zero.

Table 12 Parameter Estimates—Autoregressive Coefficients

Parameter	ϕ	σ_ϵ^2
Role composition		
Role1	0.861	0.881
Role2	0.892	0.637
Role3	0.859	0.734
Role4	0.902	0.564
Role5	0.853	0.975
Role6	0.915	0.549
Visit intensity	0.806	1.284

Note. All numbers reported are posterior means.

Table 13 Bias Correction for Search Engines and Advertising Networks

	Category percentage absolute error		
Name	Proposed model	Raw data	Improvement (%)
Advertising networks			
DoubleClick	0.545	0.965	43.52
Google adsense	0.341	0.805	57.64
Facebook social plugins	0.265	0.691	61.65
Quantcast	0.332	0.818	59.41
Search engines			
Google	0.191	0.383	50.13
Yahoo!	0.266	0.303	12.21
MSN	0.365	0.677	46.09
AOL	0.358	0.364	1.65

Table 14 Information Content of Search Engines and Advertising Networks

Name	Relative MAPE		
	Proposed model	Histogram	Improvement (%)
Advertising networks			
DoubleClick	1.662	2.213	24.92
Google adsense	1.571	2.500	37.17
Facebook social plugins	1.537	2.359	34.85
Quantcast	1.551	2.442	36.51
Search engines			
Google	1.392	1.902	26.79
Yahoo!	1.483	2.099	29.33
MSN	1.528	2.005	23.78
AOL	1.520	1.698	10.48

Notes. Relative MAPE is the ratio of MAPE for predictions made from the data available to the firm using the proposed model or simply the individual histogram, over the MAPE of using the actual profile and the proposed model, which serves as the benchmark. A smaller number indicates better performance.

firms with the most data. For firms with less data, the prior will have a stronger influence on the prediction, and hence may be more effective in correcting bias.

The predictive performance for user profiles using the data available to each firm is reported in Table 14. Performance is reported relative to that using the actual user profiles,[17] which is the performance achievable for out-of-sample prediction if the business has access to all user activities. Both predictions made with our proposed model and those directly using individual level histograms (i.e., treat the proportion of each category as observed by the firm as the true probability of visiting that category) are reported. Table 14 shows that the performance for each search engine and advertising network is worse than using the actual profile (number higher than 1 indicates more prediction errors than using the actual profile). This is as expected, as each business has only a partial view of each consumer. More importantly, for all advertising networks and search engines, the predictive performance with our proposed model is much better than that using the histogram, i.e., the raw visitation data available to each firm. This improvement in predictive performance with our model demonstrates the bias correction ability of our proposed approach at the individual consumer level, in addition to at the firm level as discussed above (see Table 13). Furthermore, comparing the different companies shows that Google has the best performance among all.[18] Interestingly, the performances of the search engines analyzed are better than those of the advertising networks. This suggests that, although search engines may cover smaller portions of consumer Web visits than those major advertising networks, their data is of higher quality. Thus, even with the smaller information set, they can quite effectively recover the consumer behavioral profile. In the RTB auction, then, search engines will have advantages over other advertising networks.

The predictive performance above is measured by comparing the predicted counts with the actual counts of each category. This shows only a part of the value of our proposed model. By incorporating heterogeneity of roles and mapping roles across categories, our proposed model would help predict not only the mean (i.e., individual level visitation counts per category) but also the distribution as a whole. To see this, we consider a common industry task of finding the most active users of a certain type. Firms often care more about the most promising leads than the rest. In our context, the task is operationalized as one wherein predictions are made on whether a specific user is among the top X% active users in each website category. This translates the task of predicting the count (i.e., a regression task) into one of predicting the tail of the distribution (i.e., a classification task). The performance of this prediction task using our proposed model and the individual consumer level histogram is reported in Table 15. For most companies, using our proposed model for classification achieves significant improvement over using the individual consumer level histogram only.[19] This is a further verification of the value of our proposed approach.

Furthermore, we investigate the predictive performance of our model along the temporal dimension, i.e., when only a limited number of days' data are available. This is of interest to firms and policy makers. Firms are constantly faced with limited information (e.g., due to cookie deletion or expiration). Policy makers must consider alternatives to limit business' possession of consumer data based on privacy concerns. For each

[17] The actual profile of a user is her complete visitation records collected by the collaborating market research firm as discussed in §4. Using the actual profile, the task is the standard out-of-sample forecasting.

[18] In our analysis, we treat each advertising network as a separate entity. In the industry, multiple advertising networks may be owned by the same firm, e.g., Google owns DoubleClick and AdSense. "Which firm has how much information" is a substantive question interesting for both policy (e.g., antitrust measures on customer information) and strategy (e.g., which other advertising network should a firm acquire to complement its own). We leave this issue for future study.

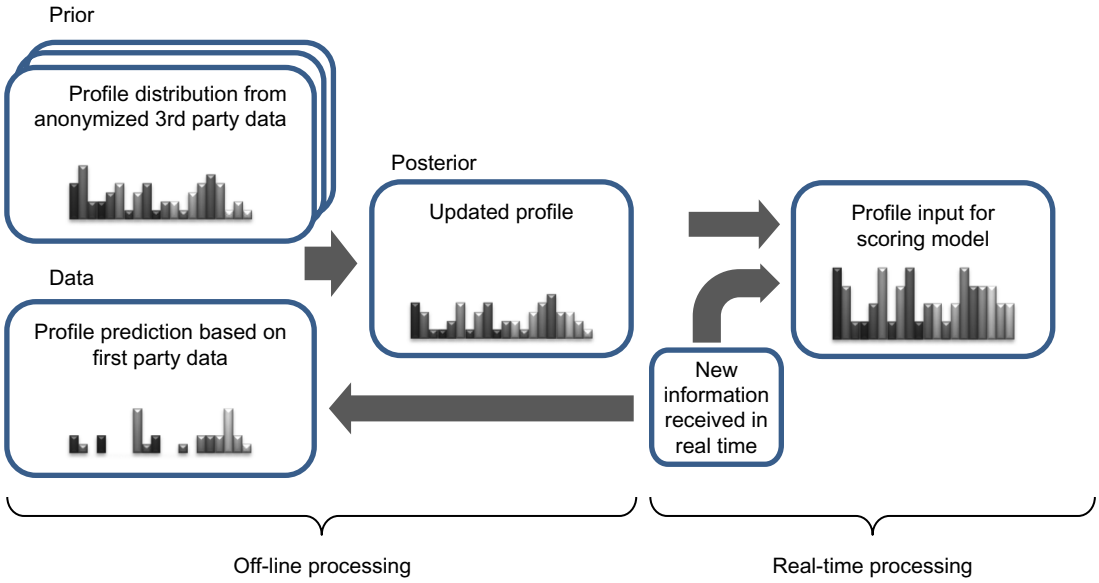
[19] For each firm, classification is performed for only those users observed by the firm (i.e., if a user's visitation is not observed by a firm, that user is excluded from the classification task). Because each firm has a different set of observed users, the prediction accuracy reported in Table 15 cannot be compared across firms. They can, however, be compared between using our proposed model and using the individual level histogram. The result reported in Table 15 is for $X = 30$. We also tested $X = 10$, 20, 40, and 50. The improvement of our proposed model increases as X increases. Details are available from the authors on request.

Table 15 Accuracy of Finding Active Users in Specific Categories

	Classification accuracy		
Name	Proposed model (%)	Histogram (%)	Improvement (%)
Advertising networks			
DoubleClick	57.53	48.24	19.26
Google adsense	51.57	39.84	29.44
Facebook social plugins	53.96	41.67	29.50
Quantcast	52.86	41.17	28.37
Search engines			
Google	56.70	47.25	19.98
Yahoo!	47.36	37.94	24.85
MSN	41.46	35.96	15.28
AOL	43.72	35.27	23.99

Notes. Performance of classifying whether each user is among top X% active users for each category. $X = 30$, Accuracy = (Precision + Recall)/2.

Figure 6 (Color online) Intuition Behind the Proposed Approach

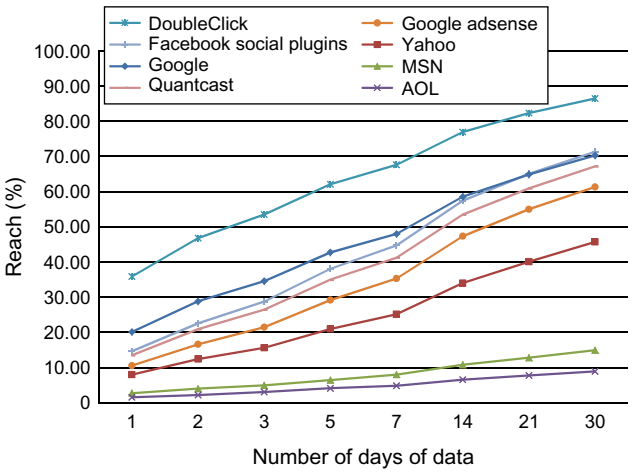


search engine and advertising network, we first calculate the increase in reach, i.e., the percentage of users who show up in the data set when more days of data become available. As Figure 7 shows, when only one day's data is available, most firms reach less than 20% of the users (i.e., they do not observe any action from the rest of the users in a day), with the exception of DoubleClick, which observes certain actions of 36% of the users.[20] As more days of data become available, the reach of all firms increases, but at different rates. Most advertising networks and search engines have a steady and significant increase over time. DoubleClick reaches 86% of users with 30 days of data, the highest among all. By comparison, however, MSN and AOL reach less than 15% of users even with a full month of data. Overall, the reach of advertising networks dominates that of search engines. Among the search engines, only Google reached more than half of the consumers with 30 days of data, while all advertising networks did so.

We evaluated classification accuracy using individual-level histograms and using our proposed model for different numbers of days of data. The improvement achieved using our proposed model compared to the individual-level histogram for each firm is plotted in Figure 8. As Figure 8 shows, when only a few days of data are available, the improvement of the proposed model over the individual histogram is even more significant than when a whole month's data is available (as reported in Table 15). By connecting multiple categories using roles, our proposed model can leverage a limited amount of information much more effectively than the individual-level histogram. As more days of data become available, the gap between an individual

[20] We used November data for this computation; the month December was used as holdout sample.

Figure 7 **(Color online) The "Reach" of Search Engines and Ad Networks**



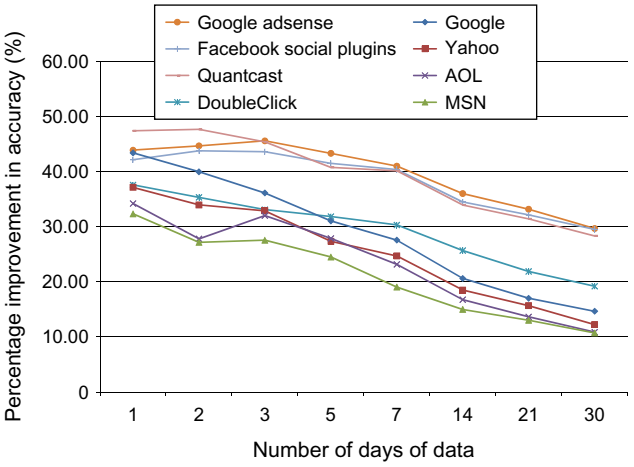
histogram and our proposed model narrows; classification using the individual histogram becomes easier. This suggests that our modeling approach is especially valuable when a limited amount of information is available. As Figure 8 also shows, search engines in general achieve more significant improvement than advertising networks.

6.2.1. Profile Prediction in Real-Time. Continuing the discussion on RTB, when new information is provided to each firm in real-time, such information can be assembled into the first-party data each firm owns. This would enhance the predictive performance of each firm. This assimilation of new information and updating of profile prediction can be done periodically offline, using the same approach discussed in §5 and used above. In an actual RTB auction, however, bid submission time is measured in milliseconds. This demands more rapid posterior inference than the MCMC algorithm discussed in §5. Our proposed model can accommodate this real-time demand using an approximation with appropriate pre-processing. This approximation is done in a two-step process. We still denote the visit information available to the firm as $\tilde{V}_i$ in §5, and denote the new information provided to the firm by the RTB's facilitating platform as $V_i^R = (v_{i1}^R, \ldots, v_{iC}^R)$. We then perform the following two steps:

- Step 1: For $c = 1, \ldots, C$, compute profile posterior $\theta_i^c \mid \{\tilde{V}_i, c\}$
- Step 2: Approximate the posterior as a weighted average: $\theta_i = \sum_{c=1}^{C} v_{ic}^R \theta_i^c / \sum_{c=1}^{C} v_{ic}^R$

In step 1, the posterior of the consumer profile is generated using the same MCMC algorithm, where the data is the known visit information $\tilde{V}_i$ plus one more visit to category c. This is performed for each category, and this step will result in C different posterior predictions. In step 2, a weighted average of these C different posteriors is calculated based on the number of visits of each category contained in the new information. The key to this approach is that only step 2 need to be performed in real time. Although step 1 likely requires more time than the few milliseconds available in RTB, it can be performed beforehand, since it does not require the real-time information V_i^R. Step 2, taking into account V_i^R, is a simple arithmetic average that can easily be done in real time. Although this weighted average is just an approximation of the true posterior, it will likely be close especially if a reasonable amount of information is already contained in $\tilde{V}_i$, and it satisfies the demand of a real-time response.

Figure 8 **(Color online) Improvement of Classification Accuracy from Individual Histogram**



6.2.2. Scenario Analysis: Display Ad Targeting. The improvement in user profiling afforded by our model may have significant profit implications for firms. In this section, we present an economic simulation that illustrates potential gains the proposed model may offer to a firm if used in individual-level targeting of display ads.

Consider a hypothetical digital advertising agency that generates traffic to a client website using display (banner) advertising. The agency distributes ads through ad exchange paying $2.80 per thousand impressions (i.e., cost per mile (CPM)) served and gets an industry average CTR of 0.5% (Johnston 2014). Accordingly, the agency's effective cost of generating a click to the client's website is $0.56. The agency charges the client a pre-negotiated rate of $0.67 per site-visit. The agency operates on a set daily budget of $1,000, which helps to generate about 1,786 visits per day with the baseline CTR of 0.5%. Clearly, the agency's profitability will improve if it can produce more clicks in the display ads campaign. While several factors contribute to the CTR of a given ad (e.g., ad creative, page placement, context), profile-based targeting is acknowledged to be one of the key drivers of ad performance (Hazan and Banfi 2013).

Table 16 Segment Sizes and Click-Through Probabilities

	Segment size (%)	Targeting premium (%)	CTR (%)
Top users	30	25	0.63
Everybody else	70	−11	0.45

As a common practice in this industry the agency uses their own proprietary scoring model that links user's online behavioral profile and propensity to click on the ad. For the sake of this simulation we assume that for the top 30% most active users in the target profile category the click-through probability is 25% higher than the average, while for the remaining 70% of users the CTR is correspondingly 11% lower than average. These numbers were selected to preserve the average rate of 0.5% (Table 16).

Using our model the agency should be able to improve the performance of this campaign by targeting individuals in the Top users segment. In the extreme case all of the ads should be served only to the Top users segment thus achieving a CTR of 0.63% (Table 16). Clearly, such a performance is unrealistic, and the effective CTR would depend on classification accuracy, which in turn depends on the information available to the agency and the targeting model. For example, assuming the agency is DoubleClick (or has information of similar quality as DoubleClick), Table 17 presents the results of effective CTR when different targeting models are used. Using the histogram approach the agency can accurately identify 42% of active users resulting in the effective CTR of 0.52%. Our model produces further improvement with 0.54% effective CTR. Finally, substituting effective CTRs from Table 17 into profit calculations we get a 25% profit improvement for the histogram-based targeting and 51% improvement for the proposed model (Table 18).

In this scenario, using our proposed model for targeting, a firm such as DoubleClick could improve profit by about 21% from the histogram-based targeting (i.e., \$297.57 versus \$246.51). A similar calculation can be made for the other firms in the data set, where the profit improvement over the base targeting model ranges from 17% (MSN) to 58% (Google). Using histogram-based targeting, profit improvement ranges from 14% (MSN) to 32% (Quantcast and Facebook Social Plugins). The economic benefit of our model for target advertising is clear and significant.

Table 17 Effective CTR Under Different Targeting Approaches

	Targeting approach (%)		
	Base	Histogram	Proposed model
Precision	30	42	54
CTR	0.50	0.52	0.54

Table 18 Profit Calculations

	Targeting approach		
	Base	Histogram	Proposed model
CTR (%)	0.50	0.52	0.54
Effective CPC (\$)	0.56	0.54	0.52
Traffic	1,786	1,860	1,937
Price to client (\$)	0.67	0.67	0.67
Revenue (\$)	1,196.43	1,246.51	1,297.57
Profit (\$)	196.43	246.51	297.57
Profit improvement over base model (%)		25.5	51.5

7. Conclusion

As digital has established itself as a key medium for reaching and interacting with consumers, one-on-one marketing is becoming a norm for online businesses. Fueling this process is the ability to collect, analyze, and act on individual-level data. This paper focuses on a fundamental component of online marketing, i.e., user profiling. User profile is a summary of a consumer's interests and preferences revealed through the user's online activity. Valued by most online businesses, user profile data has a broad application across different areas of digital marketing. McKinsey & Company regards online user profiling as one of the promising opportunities companies should take advantage of to unlock big data's potential (Hazan and Banfi 2013). However, there are a number of challenges that need to be overcome before profile data can be put to full use. First, individual consumer level records are massive. This calls for scalable high performance processing algorithms. Second, in many popular marketing applications (such as customer-based analysis) the insights generated from profile analysis must be easily interpretable to support managerial decisions. Finally and equally important, different online firms have a distinct yet limited view on their users. Hence, it is important to correct for biases that stem from firm-centric data.

Extending the CTM model (Blei and Lafferty 2007), this paper proposes a modeling approach for user profiling that aims to address the above challenges. The proposed approach augments individual-level first-party data with anonymized third-party data that significantly improves profile recovery performance and helps to correct for biases. The approach is highly scalable and easily parallelized, improving almost linearly in the number of CPUs. It produces easily interpretable and intuitive results, while taking into account observed and unobserved heterogeneities. We demonstrate the interpretive power of our model using the Web surfing data of a large online user panel, where intuitive behavior patterns and their composition (of keen interest to many online businesses) are extracted using the model.

Furthermore, we collected a unique data set of coverage areas for a variety of online businesses. Focusing

on the most powerful online businesses in terms of consumer knowledge, i.e., search engines and advertising networks, we show that although they have a significant amount of user data, their views of users' Web activities are biased. We find that while advertising networks have a higher coverage of consumer website visitations, the top search engines perform somewhat better in recovering user profiles. In other words even with relatively smaller information sets, search engines have better quality data that enable them to effectively recover a consumer behavioral profile. We also show that our modeling approach is effective for finding active users of different categories, and is especially useful for temporally limited data.

These results may have implications for policy makers as to consumer privacy regulations. Consumer online privacy is a highly debated topic (Goldfarb and Tucker 2011). The ability to compare and quantify informational advantages and disadvantages of different types of online businesses may offer an important input to regulators' decisions. For example, on a macro level, our approach can be used to explore how the informational power will be reallocated among key online firms if new regulations are introduced that limit the ability of firms to track individual activities online (on time or scope dimensions), or limit the amount or type of data that can be shared across businesses (e.g., among members of an advertising network). Our preliminary analysis shows that the impact of such regulation may vary across the firms. Another potential macro level application of our approach is in information trade. The proposed approach may assist an online firm in choosing strategic alliances or third-party data suppliers to augment first-party user information to correct for biases specific to this firm.

Going forward we see a number of ways to expand our work. Clearly, consumer information used in targeting is not limited to the behavioral profile of cross-site browsing activities. There are a number of other dimensions of user data (e.g., purchases, search terms, user-generated content, social networks, "likes," etc.) that can be of value to online marketers. While the proposed approach can accommodate such additional types of data, treating these as count data, we may develop alternative ways of profiling users on these dimensions. Exploring these approaches would be an interesting extension to our work. On the modeling front, there are a number of potential extensions that may improve the predictive performance of the proposed model.[21] First, the model can be further developed to account for possible variations in profile composition on the session level. Second, the mapping from roles to website categories can be modeled dynamically, which can further improve model flexibility. Finally, the assumption of the fixed number of roles can be relaxed by allowing variation across consumers using, for example, the Dirichlet process. Also, with our analysis we have barely scratched the surface of consumer knowledge power allocation across different online businesses. Our unique data set allows us to explore this topic in greater depth. We hope that, as one of the first papers on consumer online profiling in marketing, this research will inspire more work in this exciting field.

[21] We thank an anonymous reviewer for these helpful suggestions.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2015.0956.

Acknowledgments

The authors thank the editors, the associate editor, and two anonymous reviewers for their helpful suggestions. Funding for this work was provided by the Robert H. Smith School of Business at the University of Maryland. The authors contributed equally and are listed in reverse alphabetical order.

References

Ahmed A, Aly M, Gonzalez J, Narayanamurthy S, Smola A (2012) Scalable inference in latent variable models. *Proc. Fifth ACM Internat. Conf. Web Search Data Mining*, 123–132.

Ansari A, Mela CF (2003) E-customization. *J. Marketing Res.* 40: 131–146.

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

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

Bodapati AV (2008) Recommendation systems with purchase data. *J. Marketing Res.* XLV(February):77–93.

Danaher PJ (2007) Modeling page views across multiple websites with an application to Internet reach and frequency prediction. *Marketing Sci.* 26(3):422–437.

Danaher PJ, Smith MS (2011) Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Sci.* 30(1): 4–21.

eMarketer (2014a) U.S. mobile ad dollars shift to search apps. (June 5), http://www.emarketer.com/Article/US-Mobile-Ad-Dollars-Shift-Search-Apps/1010898.

eMarketer (2014b) U.S. programmatic ad spend tops $10 billion this year, to double by 2016. (October 16), http://www.emarketer.com/Article/US-Programmatic-Ad-Spend-Tops-10-Billion-This-Year-Double-by-2016/1011312.

Ghose A, Yang S (2009) An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Sci.* 55(10):1605–1622.

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

Goldfarb A, Tucker C (2011) Privacy regulation and online advertising. *Management Sci.* 57(1):57–71.

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

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. *Marketing Sci.* 28(2):202–224.

Hazan E, Banfi F (2013) Leveraging big data to optimize digital marketing. http://www.mckinseyonmarketingandsales.com/leveraging-big-data-to-optimize-digital-marketing.

Hofmann T (1999) Probabilistic latent semantic indexing. *Proc. 22nd Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval*, 1999.

InternetLiveStat.com (2014) Internet users by country. http://www.internetlivestats.com/internet-users-by-country/.

Jerath K, Ma L, Park Y-H (2014) Consumer click behavior at a search engine: The role of keyword popularity. *J. Marketing Res.* 51(4):480–486.

Johnson EJ, Moe WW, Fader PS, Bellman S, Lohse J (2004) On the depth and dynamics of world wide Web shopping behavior. *Management Sci.* 50(3):299–308.

Johnston M (2014) Display ad CPM rates. http://monetizepros.com/cpm-rate-guide/display/.

Lambrecht A, Tucker C (2013) When does retargeting work? Information specificity in online advertising. *J. Marketing Res.* 50(5): 561–576.

Li S, Liechty JC, Montgomery AL (2015) Modeling category viewership of Web users with multivariate count models. Working paper, Carnegie Mellon University, Pittsburgh.

Marshall J (2014) U.S. online display ad spending will nearly double by 2019. *Wall Street Journal* (October 6), http://blogs.wsj.com/cmo/2014/10/06/forrester-us-online-display-ad-spending-will-nearly-double-by-2019/.

Moe WW (2003) Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consumer Psych.* 13(1–2):29–40.

Moe WW, Fader PS (2004) Dynamic conversion behavior at e-commerce sites. *Marketing Sci.* 50(3):326–335.

Papadimitriou CH, Raghavan P, Tamaki H, Vempala S (2000) Latent semantic indexing: A probabilistic analysis. *J. Comput. System Sci.* 61(2):217–235.

Park Y-H, Fader PS (2004) Modeling browsing behavior at multiple websites. *Marketing Sci.* 23(3):280–303.

Rutz OJ, Bucklin RE (2011) From generic to branded: A model of spillover in paid search advertising. *J. Marketing Res.* 48(1):87–102.

Rutz OJ, Trusov M, Bucklin RE (2011) Modeling the indirect effect of paid search advertising: Which keywords lead to more future visits? *Marketing Sci.* 30(4):646–665.

Urban GL, Liberali G, MacDonald E, Bordley R, Hauser JR (2014) Morphing banner advertisements. *Marketing Sci.* 33(1):27–46.

Yang S, Ghose A (2010) Analyzing the relationship between organic and paid search advertising: Positive, negative or zero interdependence? *Marketing Sci.* 29(4):602–623.