THE ECONOMIC VALUE OF CLICKSTREAM DATA FROM AN ADVERTISER’S PERSPECTIVE

Nottorf, Florian and Funk, Burkhardt
Leuphana Universität Lüneburg, Scharnhorststraße 1, 21335 Lüneburg,
{nottorf, funk}@uni.leuphana.de

Abstract

With the increasing share of online advertising, the amount of highly detailed data available on an individual consumer level has substantially increased over the last few years. Because there are costs associated with collecting, analyzing, and storing this type of data, it is important to understand the value of this specific type of personal data. In this paper, we illustrate the economic value of clickstream data by highlighting the direct and interaction effects of multiple types of online advertising on an individual user-level. By developing a binary logit model with Bayesian estimation techniques to predict consumer behavior and by analyzing clickstream data from a large financial service provider, we find strong differences in the effects of repeated ad exposures. Along with uncovering important insights, such as the positive cross-channel effects of display and paid search advertising, which are important when managing online advertising campaigns, we demonstrate how the value of clickstream data can be estimated to support decision-making in the emerging context of real-time bidding (RTB).

keywords: personal data valuation, clickstream analysis, online advertising efficiency, consumer behavior, Bayesian estimation
1 Introduction

By the 1960s and 1970s, the value of personal data had already been recognized. Since then, it has been a matter of debate whether data privacy leads to market inefficiencies as was argued by the Chicago School (Posner, 1978, 1981) or whether it might increase economic welfare (Hermalin and Katz, 2006). However, it is clear that the benefits and costs of disclosing and collecting personal data differ between the two major stakeholders involved: (i) data subjects, who protect or disclose their data, and (ii) data holders, who collect, analyze, and leverage personal data for their economic activities (Acquisti, 2010). While data subjects disclosing personal data might profit from targeted offerings, be it products, services, or personalized information (Varian, 1996), the potential advantages for the data holders encompass increased revenues through price discrimination using personal data (Acquisti and Varian, 2005) and efficiency gains (e.g., increased advertising efficiency). In this paper, we focus on the latter and show how a company investing in online advertising can estimate the value of clickstream data.

In the last decade, the options for online advertising have become increasingly complex, leading to the necessity of making sophisticated decisions. At the same time, new technologies offer companies the potential to “follow” individual users across multiple types of online advertising on a highly detailed, individual consumer level (Kauffman et al., 2012; Lambrecht and Tucker, 2011). These clickstream data are collected through cookie tracking by companies’ advertising servers and is associated with substantial costs. Both the financial investments and the availability of personal data have awakened new interest in correctly attributing online advertising success to multiple types of online advertising.

Companies doing online advertising often use standardized metrics and simple heuristics to evaluate the profitability of each specific online advertising campaign. However, these ratios do not capture the consumers’ decision-making processes over time and do not account for the interaction effects between multiple advertising activities. Using clickstream data while accounting for the cross-channel effects of online advertisement exposures may reveal important economic insights for advertising agencies (i.e., relocate advertising specific spending) and increase efficiency in regard to real-time bidding (RTB).

Although there have been attempts to attribute success to multiple types of online advertising (Dalessandro et al., 2012b; Shao and Li, 2011), the effect of advertising on the individual behavior of consumers has not yet been comprehensively analyzed. Therefore, with this work, we can contribute to the existing knowledge in several ways. Specifically, we analyze whether the costs of tracking and analyzing user-journeys—such as consumers’ touch-points (impressions) across multiple types of advertising—are economically reasonable. We develop a binary logit model using Bayesian estimation techniques that accounts for the direct and interaction effects of online advertising on consumer click probabilities. Although the “click” is just one of multiple options to evaluate online advertising activities (Dalessandro et al., 2012a), that click is still a mandatory step for a channel-specific conversion that reveals important insights into the effectiveness of current and past advertising activities.

This remainder of the paper is structured as follows: first, we briefly review the previous research related to this work. Second, we specify how to model the click probabilities of display ads and simultaneously account for multiple online advertising efforts. The model is then applied to empirical data from a large financial service provider. After outlining our findings and discussing our results, we conclude this work by highlighting its limitations and providing suggestions for future research.

2 Related Work

The click-through rates (CTR) of banner advertising decreased from 2% in 1995 (Cho and Cheon, 2004) to 0.3% in 2004 to 0.08% in 2010 (Chapman, 2011). Repeated exposure to advertising is often considered
annoying and less effective with regard to click-through rates (Cho and Cheon, 2004; Edwards et al., 2002; Li et al., 2002). Nonetheless, Manchanda et al. (2006) and Yoon and Lee (2007) show that banner advertising intrinsically enhances consumer clicking and purchasing behavior. Their findings imply that the CTR may be an inadequate indicator for analyzing the effectiveness of banner ads. Dallessandro et al. (2012a) confirm this and demonstrate that considering only the click may mislead managers when evaluating online advertising campaigns. The authors show that there are alternative proxies to evaluate and optimize online advertising campaigns (such as onsite clicks and time on a company’s website). New types of display advertising, such as contextually targeted (Goldfarb and Tucker, 2011) or emotionally personalized ads (Meyer et al., 2011), add further complexity and require additional advertising measures (Lambrecht and Tucker, 2011).

In recent years, paid search advertising has become the favored tool in online advertising because it produces higher CTRs compared with banner advertising (Ghose and Yang, 2009). An emerging stream of research is analyzing paid search advertising from the advertiser’s perspective and providing important insights into consumer clicking and purchasing behavior (see, for instance, Rutz et al. (2012) and Nottorf and Funk (2013)). Although the number of studies on paid search advertising has increased significantly, there have been few attempts to examine an integrated model of online advertising efforts (Dinner et al., 2011; Wiesel et al., 2011) or to standardize attribution for online advertising as an alternative from the “last cookie” wins attributions (Dallessandro et al., 2012b; Shao and Li, 2011). For example, there is just one work that investigates the effects of the interaction between display and paid search advertising on an individual user level (Nottorf, 2013).

3 The Valuation of User-Level Data Across Multiple Online Advertising Channels

3.1 Model Specification

In our approach, we account for the consumers’ user-journey and the effects of exposure to multiple display advertisements and paid search advertising to model the consumers’ individual click probabilities. We interpret all of the ads that a consumer is exposed to over time as a repeated number of discrete choices (Bucklin and Sismeiro, 2003). For example, consumers can decide whether to obtain additional information regarding a product after each exposure to each display ad, which results in a click/non-click decision. Note that we model the consumer choice of clicking or not clicking on a display ad (binary choice) by incorporating the effects of exposure to repeated display ads, such as banner or video ads, and paid search ads, as explanatory variables. We consider short-term advertising effects on consumers’ click probabilities by adding variables to the model specification that vary across time \( t \) \((X_{ist})\) as well as their long-term effects by incorporating variables that only vary across sessions \( s \) \((Y_{is})\). To model the individual contribution of each advertising effort and its effect on the probability that a consumer \( i \) will click on a display ad, we specify a binary logit choice model that extends the specification of Chatterjee et al. (2003). The probability that consumer \( i \) clicks on a display ad at time \( t \) in session \( s \) is modeled as follows:

\[
Click_{ist} = \begin{cases} 
1 & \text{if user } i \text{ clicks on a display ad at time } t \text{ in session } s \\
0 & \text{otherwise,} \end{cases}
\]  

(1)

with the probability

\[
Pr(Click_{ist} = 1) = \frac{\exp(1_{ist} \alpha_i + X_{ist} \beta_i + Y_{is} \gamma_i + \varepsilon_{ist})}{1 + \exp(1_{ist} \alpha_i + X_{ist} \beta_i + Y_{is} \gamma_i + \varepsilon_{ist})},
\]  

(2)
The variables clicked on an ad, the variable remains zero. Thus, \( I_{ist} \) is an indicator function that assumes the value 1 if a consumer clicked on the previous display ad in session \( s \). All three variables account for the propensity of an individual consumer to click on a banner ad, retargeted ad, or video ad, independent of the total number of ads that the consumer has been exposed to.¹ The prior research states that consumer responses to banner ads are highly dependent on individual involvement (Cho, 2003; Danaher and Mullarkey, 2003) and exhibit strong heterogeneity (Chatterjee et al., 2003).

After introducing our general model specification, we specify the covariates in greater detail:

\[
I_{ist} = \{I_{ist}^{\text{ban}}, I_{ist}^{\text{ret}}, I_{ist}^{\text{vid}}\},
\]

where \( I_{ist}^{\text{ban}}, I_{ist}^{\text{ret}}, \) and \( I_{ist}^{\text{vid}} \) are indicator functions that assume the value of 1 if consumer \( i \) is exposed to the respective advertising type at time \( t \) in session \( s \). All three variables account for the propensity of an individual consumer to click on a banner ad, retargeted ad, or video ad, independent of the total number of ads that the consumer has been exposed to.¹

We expect the effect of repeated ad exposure to vary depending on the type of display ad that is repeated and whether the consumer has previously clicked on a paid search ad. We generally expect stronger wear-in effects² from the advertising activities after the consumers have searched and clicked on a paid search ad—paid search ads appear in response to a consumer’s own queries and are far less intrusive and more appealing to consumers—this interaction may therefore positively increase the subsequent effects of repeated exposure to display ads. This assumption is based on previous research by Danaher and Mullarkey (2003), who found that online advertising is more effective for consumers in a goal-directed mode.

Thus, \( x_{ist}^{\text{ban}}, x_{ist}^{\text{ret}}, \) and \( x_{ist}^{\text{vid}} \) refer to the cumulative number of banner, retargeted, and video ads to which consumer \( i \) has been exposed until time \( t \) in session \( s \) if that consumer has not previously clicked on a paid search ad. \( x_{ist}^{\text{ban}|\text{sea}} \) and \( x_{ist}^{\text{ret}|\text{sea}} \) also represent the cumulative exposure to banner and retargeted ads within session \( s \) until time \( t \) but are nonzero only if consumer \( i \) has previously clicked on a paid search ad.³ \( x_{ist(t-1)}^{\text{click}} \) is the cumulative number of clicks on display ads until \( t - 1 \) in the current session \( s \). \( \text{Click}_{ist(t-1)} \) is an indicator function that assumes the value 1 if a consumer clicked on the previous display ad in \( t - 1 \). \( \text{TLClick}_{ist} \) refers to the logarithm of time since a consumer last clicked on an ad. If a consumer has not clicked on an ad, the variable remains zero.

The variables \( Y_{ist} \), which vary across sessions and consumers, are specified as follows:

\[
Y_{ist} = \{Y_{ist}^{\text{ban}}, Y_{ist}^{\text{ret}}, Y_{ist}^{\text{vid}}, Y_{ist}^{\text{ban}|\text{sea}}, Y_{ist}^{\text{ret}|\text{sea}}, Y_{ist}^{\text{click}}, Y_{ist}^{\text{ist}|\text{Session}_{ist}}, Y_{ist}^{\text{ist}|\text{Session}_{ist}}\}.
\]

¹We define banner ads as nonspecific “generic” display ads that are delivered to consumers. Retargeted (banner) ads are delivered to “known” consumers. Video ads are multimedia ads that are displayed on YouTube. See section “User-Level Data” for further details.

²Exposures to ads are believed to wear in at a particular level of repetition if they have a significant positive effect on consumers (Pechmann and Stewart, 1988).

³Please note that we also incorporated \( x_{ist}^{\text{vid}|\text{sea}} \) in the model specification but omitted this variable from the final specification because of a lack of data.
\( y_{\text{ban|sea}}^{\text{click}} \) and \( y_{\text{ret|sea}}^{\text{click}} \) capture the long-term effects of banner and retargeted ad exposures, respectively, if a consumer previously clicked on at least one paid search ad. \( \gamma_1^{\text{click}} \) is the cumulative number of clicks on display ads in the previous sessions. IST \( t_s \) is the logarithm of the intersession duration between session \( s \) and \( s-1 \) and remains zero if a consumer is active in only one session. Session \( s \) refers to the number of sessions during which a consumer has already been exposed to ads.

### 3.2 User-Level Data

We use a unique dataset from a large European financial service provider that wants to remain anonymous. The company operates several display and paid search advertising campaigns to advertise products such as giro accounts and instant access savings accounts. User-level data that follows consumers across multiple advertising channels and types is collected through cookie-tracking. This tracking enables us to distinguish the following types of advertisements on an individual consumer-level. First, we distinguish (normal) banner ads, which are nonspecific “generic” ads that are delivered to consumers. The second type are retargeted (banner) ads, which are delivered to “known” consumers. For all types of banner ads, the financial service provider advertises its product in the same manner. Third, the video ads in this study are multimedia ads that are displayed on YouTube. Furthermore, the dataset contains information on consumer clicks on paid search ads. Because we model only a consumer’s direct response to display ad exposures (banner, retargeted, and video), we consider the consumer activities on search engines that occur chronologically between display ad exposures (see \( x_{\text{ban|sea}} \) and \( x_{\text{ret|sea}} \)).

Please further note that due to a lack of data, we are not able to control for past conversions. Although there might be consumers who purchased something from the company in question and might be exposed to display advertisements afterwards, the effect of past conversions can be neglected to some extent because the number of users who purchased and later were exposed to display advertisements is very small. Especially in the financial sector investigated here (where a conversion is a very rare event), including information on past conversions would not significantly affect the parameter estimates.

The dataset contains a random sample of 150,000 users from a 48-day period (November 27, 2011 to January 13, 2012) and, therefore, only represents a small cohort from a short investigation period. 

---

4These retargeted ads are served by the Google Display Network on the basis of a cookie that is stored on a consumer’s personal computer (i.e., because of the previous browsing history for financial products or previous visits to the website of the company in question).

5Because of technical restrictions, there is no information on individual searches by consumers that do not include clicking on a paid search ad.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean (95% cov. interval)</th>
<th>Variables</th>
<th>Mean (95% cov. interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{ist}$</td>
<td>-9.13 (-10.33, -8.33)</td>
<td>$Y_{ist}$</td>
<td>-1.88 (-2.60, -1.58)</td>
</tr>
<tr>
<td>$x_{ban}^{ist}$</td>
<td>-110.15 (-116.54, -106.24)</td>
<td>$y_{ban}^{ist}$</td>
<td>-2.33 (-2.66, -1.82)</td>
</tr>
<tr>
<td>$x_{vid}^{ist}$</td>
<td>-2.92 (-2.51, -2.30)</td>
<td>$y_{vid}^{ist}$</td>
<td>-1.63 (-1.88, -1.44)</td>
</tr>
<tr>
<td>$x_{ban</td>
<td>sea}^{ist}$</td>
<td>1.12 (0.61, 1.99)</td>
<td>$y_{ban</td>
</tr>
<tr>
<td>$x_{ret</td>
<td>sea}^{ist}$</td>
<td>-4.59 (-4.97, -4.19)</td>
<td>$y_{ret</td>
</tr>
<tr>
<td>$x_{click}^{ist(-1)}$</td>
<td>15.71 (13.85, 18.41)</td>
<td>$y_{click}^{ist(-1)}$</td>
<td>7.88 (7.13, 9.07)</td>
</tr>
<tr>
<td>$Click_{ist(-1)}$</td>
<td>9.84 (9.42, 10.70)</td>
<td>$IST_{ist}$</td>
<td>5.17 (4.71, 5.68)</td>
</tr>
<tr>
<td>$TLClick_{ist}$</td>
<td>5.08 (3.51, 6.78)</td>
<td>$Session_{ist}$</td>
<td>0.99 (0.71, 1.57)</td>
</tr>
<tr>
<td>$I_{ist}$</td>
<td>-24.87 (-27.09, -22.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$I_{ban}^{ist}$</td>
<td>-28.32 (-30.50, -26.04)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$I_{vid}^{ist}$</td>
<td>-31.21 (-34.39, -29.91)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Parameter estimates of the proposed model. Notes: We report the mean and the 95% coverage interval; significant estimates are in boldface.

4 Results

Similar to Chatterjee et al. (2003), we use a Bayesian standard normal model approach to account for consumer heterogeneity and to determine the set of individual parameters. We apply an MCMC algorithm including a hybrid Gibbs Sampler with a random walk Metropolis step for the coefficients for each consumer and utilize the R-package bayesm by Rossi et al. (2005). We perform 60,000 iterations and use every twentieth draw of the last 20,000 iterations to compute the conditional distributions.

As was argued in chapter 2, the results of the analysis (see Table 1) show a very low click tendency (see $I_{ist}$). The logits of these estimates indicate a click probability that ranges from $1.6 \times 10^{-11}$ to $2.8 \times 10^{-14}$. Apart from this finding, it appears that the consumers either click on one of the first ad exposures or are unlikely to ever click on an ad within a discrete session, as the repeated exposures to display ads have a significant negative effect on the click probability of consumers (see $x_{vid}^{ist}$). In particular, the repeated exposure to retargeted ads within a current session strongly reduces the consumers’ click probability (see $x_{ret|sea}^{ist}$). Our results indicate that the expected click probability of a consumer is the highest for the first ad exposure and subsequently decreases. Note that this finding does not apply to the repeated banner effects to which consumers are exposed after they have clicked on a paid search ad (see $x_{ban|sea}^{ist}$). This positive effect (from previous clicks on paid search ads) on the effectiveness of banner and retargeted ad exposures was not discovered previously and is consistent with our prior expectations: the previous interaction with a paid search ad (i.e., searching for related terms and clicking on a company ad) increases the effect of downstream display ad exposures.
In contrast to the prior findings of Chatterjee et al. (2003) from ten years ago, the long-term effects of repeated display ad exposure are negative (see $y_{it}^{\text{click}}$). As expected, the number of previous clicks by consumers positively influences their click probability in the next ad exposure (see $x_{it}^{\text{click}}$, $\text{Click}_{it(-1)}$, and $y_{it(t-1)}^{\text{click}}$). The consumers who had already clicked on a link demonstrated a higher probability of clicking again than the consumers who clicked less or who had not yet clicked at all. Those consumers who had clicked can be regarded as “locked-in” consumers because they have either already been successfully attracted by the services and products of the company in question or have already actively noticed the ads, compared with consumers who have not yet clicked.

5 The Valuation of Personal Data

We have successfully demonstrated how to model consumer clickstreams. Next, we derive and discuss the economic value of revealing insights regarding the cross-channel advertising effects and the value of having and analyzing the available clickstream data in the context of real-time bidding (RTB).

5.1 Cross-Channel Insights

The availability of clickstream data offers companies doing online advertising the ability to understand the consumer process of repeated display advertising exposure and to more precisely account for the specific effects from the ads on consumer behavior. These findings are extremely helpful when managing and evaluating online advertising campaigns. For example, while we find that retargeted banner advertisements appear to be perceived as obtrusive in the short term, their effects do not strongly differ from the normal banner ads in the long term. In contrast to banner or retargeted ad exposures, video advertisements do not appear to be perceived as particularly obtrusive: we find that consumers’ repeated exposures to rich media advertisements such as videos are helpful in building and increasing brand awareness (i.e., in branding campaigns) because they only barely negatively affect consumer clicking behavior. Furthermore, we investigated the predominantly positive interaction effects from paid search ad exposures on the effects of exposure to repeated display ads. While we provide important insights on the direct as well as the interaction effects of online advertising on consumer behavior, we leave the question of a correct attribution for cross-channel advertising activities on the companies’ overall success measures open for future research.

5.2 Real-Time Bidding

To demonstrate the economic value of having and analyzing user-level data in the context of RTB, we first highlight the out-of-sample fit performance of our proposed model. Therefore, we predict the actual outcome for each consumer for the next period (click/no click) and compare them with the actual, observed choices. As reported in Table 2, the prediction for the consumers’ actual choices in $t + 1$ is quite good (e.g., 43.7% of the clicks can be forecast correctly.). However, for the companies using RTB, it may not only be important to separate the consumer into two types (click/no click) but also to differentiate their actual bidding behavior (e.g., maximum bid for a display impression) along the consumers’ individual success rates (in our case: consumers’ individual click probabilities).

---

6In RTB, display ad impressions are bought in an auction-based process and displayed in real time on the individual consumer level. That is, the knowledge of a consumer’s success probability (such as a click or a conversion) at any given time is vital for accurately evaluating each ad type and appropriately adjusting financial resources.
Let us now assume the position of an advertiser that engages in a RTB setting. Depending on the individual setting (i.e., contribution margin of the advertised product), companies usually determine a specific maximum amount of money that they are willing to spend to acquire new customers (cost per order - CPO). In the following example, we assume a constant conversion rate for reasons of simplicity and thus can translate a maximum CPO to a maximum cost per click (CPC) that companies will not outbid when using RTB.

We demonstrate the economic value of clickstream data by running a second out-of-sample test. We rank order all consumers by their individual click probabilities at time $t+1$, separate them into quintiles, check how many clicks each of the quintiles actually receives in $t+1$, and calculate their respective CTR (see Table 3). For example, the quintile with the lowest 20% of click probabilities (0-20%) receives 1.5% of the total clicks in the next period, whereas 20% of the consumers with the highest click probability (80-100%) receive nearly 75% of the clicks. Concentrating the company’s bidding behavior on these upper quintiles bins may uncover potential savings.

For demonstration purposes, we simply assume a cost per mile (CPM) for display advertising in $t+1$ (in RTB, the CPM might vary from impression to impression, but this would only marginally change the following procedure for estimating the value of clickstream data). Based on the expected CTR, we are able to calculate the expected CPC. We do this for two scenarios with different CPM for each quintile (see the right hand side of Table 2). For the sake of clarity, let us consider an example and assume a maximum CPC of 0.25 €. Given that maximum, we see that in scenario 1, 20% of consumers with respect to the expected CPCs should be exposed to a display ad because the CPC of all of the other consumers is expected to be higher than 0.25 €. A company that does not have information on the clickstream data would have exposed every consumer to display ads in the first scenario because the company would not have categorized consumers along their individual click probabilities and, with 0.15 €, the total expected CPC is less than the maximum CPC. In the second scenario with an increased CPM, the company would not expose any consumer, although the consumers with the highest expected CTR should have been exposed to display ads.

The procedure outlined above leads to additional profit ($profit_{add}$) in contrast to a company that does not hold and analyze consumer-level data. To illustrate this result for scenario 1, we must consider the opportunity cost of a “lost” click ($cost_{opp}$) of a consumer (that we do not expose to display ads because we focus on the consumers with the topmost click probabilities) times the number of lost clicks ($clicks_{lost}$). Simultaneously, we save on the consumers ($user_{lost}$) that we do not expose to display advertising due to an expected CPC that is too high:

$$profit_{add} = user_{lost} \cdot \frac{CPM}{1000} - clicks_{lost} \cdot cost_{opp}$$

We assume the cost of a lost click to be equal to the maximum CPC (0.25 €). Given that assumption, the expected profit is 3.91 € for scenario 1. In the second scenario, a company that does not use the

<table>
<thead>
<tr>
<th>Observed Choice</th>
<th>Click</th>
<th>No Click</th>
</tr>
</thead>
<tbody>
<tr>
<td>Click (71)</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td>No click (4,807)</td>
<td>32</td>
<td>4,775</td>
</tr>
<tr>
<td>Total (4,878)</td>
<td>63</td>
<td>4,815</td>
</tr>
</tbody>
</table>

Table 2: Out-of-sample fit performance for $t+1$. 

---

7That is, the model was estimated on the basis of $t$ impressions per consumer.

8We do so following Chatterjee et al. (2003) with respect to Morrison (1969), who suggested to ranking observations in decreasing order to their predicted probabilities and classifying the first $x$ as clicks (where $x$ is the total number of clicks observed in the holdout sample) because the behavior to be predicted is relatively rare and the base probability of the outcome is very low. As Chatterjee et al. also point out, with a large number of nonevents (no clicks) and very few events (clicks), logistic regression models can sharply underestimate the probability of the occurrence of events.

9In a real setting, these expected CPCs should be calculated repeatedly: first, because the parameter estimates may change over time and second, to analyze the probabilities of new consumers.

10Note that there are additional costs (i.e., costs of data storage or costs for the analysis of consumer-level data), which also
information derived from clickstreams would lose 8.89 € because it misses 20% of the consumers with the highest predicted probabilities.\textsuperscript{11} Please note that this profit/loss applies only to the small sample of 4,878 consumers. In a real world scenario, a company that engages in RTB makes millions of decisions and may save thousands of dollars each day. For example, in the first scenario and given only 1 million decisions, the company would save 801 € by classifying users along the predicted click probabilities.

### 5.3 Limitations

Although we analyzed a large dataset at a highly detailed, individual user-level, these clickstream data are just one small excerpt from consumers’ daily surf routines. The user-journey that we analyzed only consists of touchpoints with the company in question. That is, there is no user-level information if the company in question has not exposed the user to a display ad (or, e.g., a 1 × 1 view-pixel). Another deficiency in our research methodology and its underlying data is that we only take into account a “click” as the indicator for the success or failure of the advertising campaign (without any information on consumer conversions). As noted earlier, the click is one but not the only approximation for an economic evaluation of online advertising activities (Dalessandro et al., 2012a). For example, in branding campaigns, the focus lies in creating and increasing brand awareness; direct responses to online advertisements (i.e., clicks or conversions) are not the primary concern.

Furthermore, there are certain limitations to tracking consumers through the application of cookies that should not be overlooked. For example, we do not have combined information for the consumers who use web browsers across multiple devices (i.e., personal computers at work versus at home), and thus we are unable to model complete sessions for all consumers. That is, we handle each cookie on a personal computer as one consumer and do not consider the use of the same personal computer by

\[\text{loss}_{\exp} = 4878 \times 0.2 \times \frac{4.21}{1000} - 52 \times 0.25 \text{ €} \]
multiple consumers (i.e., families) or the switching between workstations or browsers by consumers across multiple devices. Furthermore, current web browsers (i.e., Internet Explorer, Firefox, Chrome, or Safari) allow the usage of particular routines to delete cookies from personal computers (i.e., after each session or every day).

In our modeling approach, each parameter estimate linearly adds to another, which might not realistically describe click probabilities across all consumers. For example, there may be some users who have seen many retargeted display ads but still click on a display ad; within our additive logit approach, that consumer would have a click probability of nearly 0. These extreme cases may be better explained by the use of an exponential approach. We leave this interesting question open for further research.

Estimating the economic value of consumer-level data, the context of RTB only provides a lower limit to the true value of personal data from the perspective of the data holder. For example, there was no additional information available (such as internal, customer-relationship data or consumer-specific data such as income, interests, and age) which might have further improved the accuracy of the model’s estimation. Nonetheless, together with uncovering important insights into the cross-channel effects of display and paid search advertising as well as their different effects on individual consumer behavior, we demonstrated a variety of economic advantages that come together with clickstream data.

6 Conclusion

There are potential advantages for data holders that encompass increased revenues through price discrimination and efficiency gains using personal data (Acquisti and Varian, 2005). In this paper, we focus on the latter and derive the economic value of collecting, storing, and analyzing consumer-level data from the online advertising perspective. We developed a binary logit model with a Bayesian normal prior approach that models consumer clickstreams across multiple online advertising channels and considers the effects of the interaction between several types of online advertising.

We found that the availability (collecting and storing) and analysis of consumer-level data makes good economic sense for multiple reasons. First, the identification of individual consumer behavior is essential for understanding and evaluating multiple online advertising channels simultaneously. To be specific, our results indicate that repeated exposure to display ads decreases the consumers’ probability of clicking. The results further show that the repeated effects of display advertising strongly differ depending on which type of display ad is being exposed to a consumer. For example, the repeated exposure to video ads only barely negatively affects the consumers’ click probability (that is, the probability of the consumer choosing to click on an ad is nearly identical for each video ad exposure), but the click probability significantly decreases when the consumers are repeatedly exposed to banner ads. The most negative effect is observed for repeated exposures to retargeted (banner) ads. The effect of the repeated display ad exposures within a current session becomes less negative (retargeting) or even positive (banner) if the consumers previously searched for keywords related to the product that the company advertises and clicked on a paid search ad. This positive effect—from previous clicks on paid search ads—on the effectiveness of banner and retargeted ad exposures was not discovered previously, but is consistent with the prior research that finds that online advertising is more effective for consumers in a goal-directed mode (Danaher and Mullarkey, 2003).

Second we ran two out-of-sample tests to demonstrate fit performance. In the first test, we successfully predicted consumer clicks with our proposed model (43.7% of the clicks in \( t + 1 \) were forecast correctly). In the second test, we rank ordered all consumers by their individual predicted click probabilities and calculated CTRs and CPCs for two different scenarios. We further derived significant economic profits from having and analyzing clickstream data. The outlined procedure appears to be especially applicable
in regard to real-time bidding (RTB), where ads are bought and displayed in real time on an individual consumer level.

Although we provide a new perspective on consumer online behavior and on how online advertising influences this behavior, there is still room for further analyses. First, the integration of information on consumer conversions could uncover interesting insights because the conversion probabilities might also differ across several types of online advertising. Second, consumer-specific information, such as gender or interests, would further enrich the analysis of individual click and conversion probabilities. Finally, we also see the application of a field experiment as very important for practitioners, in which the findings from the analysis of the consumer’s individual click and conversion probabilities are applied and compared to approaches that are already used in companies’ daily activities.

References

Acquisti, A. (2010). The economics of personal data and the economics of privacy.


Dalessandro, B., Hook, R., Perlich, C., and Provost, F. J. (2012a). Evaluating and optimizing online advertising: Forget the click, but there are good proxies.


