Uncovering Bias in Ad Feedback Data Analyses & Applications*

Marc Bron
Oracle

Ke Zhou
University of Nottingham

Andy Haines
Apple

Mounia Lalmas
Spotify

ABSTRACT

Electronic publishers and other web-companies are starting to collect user feedback on ads with the aim of using this signal to maintain the quality of ads shown on their sites. However, users are not randomly sampled to provide feedback on ads, but targeted. Furthermore some users who provide feedback may be prone to dislike ads more than the general user. This raises questions about the reliability of ad feedback as a signal for measuring ad quality and whether it can be used in ad ranking. In this paper we start by gaining insights to such signals by analyzing the feedback event logs attributed to users of a popular mobile news app. We then propose a model to reduce potential biases in ad feedback data. Finally, we conclude by comparing the effectiveness of reducing the bias in ad feedback data using existing ad ranking methods along with a new and novel approach we propose that takes revenue considerations into account.

1 INTRODUCTION

In today's society we often rely on free web services for many of our daily activities such as emailing, consuming digital content and social networking. One revenue model, prevalent among Internet companies providing such free services, revolves around sponsored posts or advertisements (ads for short). In this model, advertisers pay Internet companies to show ads to their users in a way that some will engage with the advertised message and thus deliver a return on investment to advertisers. Because of the self service nature and scale of these advertising platforms, a delicate balance needs to be struck between providing a rewarding investment to advertisers, whilst at the same time minimizing negative impact to users through caused by "bad" ads, or ads of poor quality that creep into the system.

In recent years, research has focused on using historical event logs or editorial judgements to develop implicit methods that ensure users are served relevant [2, 10, 11, 14, 18, 19, 26] and high quality ads [5, 30] and to define the impact and importance of both criteria [13, 15, 21]. These methods, however, are not perfect and some online companies have incorporated feedback mechanisms that allow users to express their opinion about ads explicitly. Ad feedback tools are used by companies like, Facebook, Twitter, Yahoo and YouTube. Such signals have the potential of improving the performance of existing quality models and deriving new ones.

Before applying feedback data to such scenarios, there are two important challenges to overcome. First, we must ask whether ad feedback as a signal is generalizable beyond individual users and representative of a service's user base. The fact that ads are targeted introduces the potential for the presence of bias in ad feedback signals. Figure 1 illustrate this. Moreover, users may not all respond equally to ads as some are more prone to provide feedback than others. Second, experimentation with ad quality signals, such as ad feedback, comes at a cost in terms of the short-term revenue of Internet companies [15]. Especially online services operating solely under the sponsored ad business model may be unwilling to experiment with ad quality signals to improve their users' experience when that may have a major revenue impact, e.g., a 50% reduction in ad impressions [15]. To allow online services keeping improving quality their is a need to be able to experiment with ad quality signals in a commercial setting while managing the potential revenue impact.

To address these challenges we start with an analysis of a large dataset of ad feedback data captured from the ad feedback mechanism of a large Internet media company. This mechanism allows users to provide feedback by hiding certain ads they are exposed to. We investigate to what extent the association between ads and ad feedback is affected by the fact that ads are targeted at users with particular demographics, interests, and behaviour. Furthermore, since user may differ in terms of their responses to ads we analyse to which extent user behaviour (e.g. clicks) affects feedback.

With this understanding we then start to answer the question as to where and how *bias* resides in feedback data as well as how feedback data may be incorporated in a commercial ad serving setting, while accounting for this bias. We develop a model that *corrects* for the bias in ad feedback data to produce ad quality scores that reflect the quality of ads in their truest sense. We then introduce a technique that provides explicit control over the revenue impact when incorporating ad quality scores in the ad ranking mechanism of an online ad auction. Finally, with all this in place we compare two models, that provide biased and bias corrected feedback estimates, to investigate the value of correcting for bias in ad feedback data using various ad ranking strategies.

2 RELATED WORK AND MOTIVATION

Research in online advertising have mostly focused on predicting how an ad will perform e.g. [4, 6, 23, 24, 26]. The performance of an ad campaign is measured using a *score*, usually *click-through rate* (CTR), which is the number of times the ad was clicked out of the number of times it has been shown (number of ad impressions). Predicting CTR has been studied for many types of ad formats, e.g. [2, 4, 10, 11, 14, 18, 19, 22, 24, 26, 28, 29]. Additionally, in sponsored search, accounting for the relationship between the ad landing page, the query [6, 11, 19, 26] and the (dwell) time a user spent on the ad landing page [14] have been shown to be beneficial.

Each impression a user makes by visiting a publisher site the ad platform services is determined by an auction where the ad with the highest expected revenue wins and is thus shown to the user. Given a request for an ad to be served for that impression, ads are ranked accordingly through a real-time auction mechanism e.g. [3, 17], combining the bid - the amount of money the advertiser is willing to pay for its ad to be shown - and the score. This is formalised as

*This work was carried while all authors were at Yahoo Labs.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317304

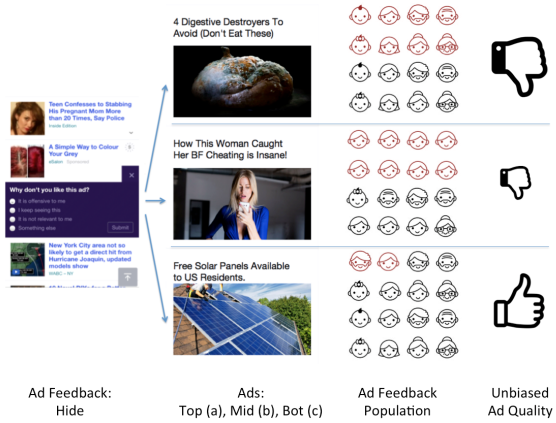


Figure 1: An anecdotal example of ad feedback vs. ad quality.

$bid \times score$, where the latter usually corresponds to the predicted probability of the ad being clicked.

The publisher on the other hand may also want to restrict ads that could have detrimental effects to the long-term engagement of their users even if it means taking a potential hit in short term revenue. Whilst ads for the most part aim to provide users with useful information about products, at the same time they may annoy, distract or offend users.

In display advertising, [13] showed that "annoying ads have a real cost to users beyond mere annoyance", such that users develop negative attitudes toward host sites, causing reduced visits of shorter duration, fewer referrals, and overall *long-term* user disengagement. Ad features that related most to ad annoyance, include animation (e.g., motion), attentional impact (e.g., distracting), aesthetics (e.g., ugly), reputation (e.g., spam, fake), and logic (e.g., confusing, unfocused). In sponsored search, [15] showed the benefit of presenting users high quality ads (e.g. ad load time on mobile) on long-term business impact. Finally, regarding the negative effects of ads, [7] made important recommendations for both publishers and advertisers, with particular attention to the rise of ad blockers - a serious threat to publishers whose revenue largely relies on advertising.

In advertising, an important criteria of quality is how relevant the ads are. In sponsored search, approaches incorporating relevance in deciding which ads to return have been effective [11, 14]; in addition, [8] showed that it is better to not show any ads than to show non-relevant ones. Some works have focused on building models of ad quality, allowing to predict the probability for an ad to be bad, and to use the predicted probability in the estimation of the ad score, e.g. [5, 12, 21, 30]. For example, ads with very short time spent on their landing pages (high bounce rate) have been shown to be detrimental to long-term engagement, and when ranking ads with an ad quality component led to higher CTR [5, 21]. The message of the work discussed above is clear - returning ads of whatever definition of poor quality you chose results in negative consequences to long term revenue. It is for this reason that using explicit feedback mechanisms from users can help capture all these effects and once integrated directly into the ad ranking score allows ads to be ranked in terms of both *short-term* and *long-term* expected revenue.

Closest to our work is [30] who used ad feedback to train models to predict bad quality ads. Using the ratio of times an ad was marked offensive over the number of times it was shown to users (defined as "offensive rate") as an objective function they found the top predictive features to be trustworthiness, the product/service provided, the brand and layout, aligning with those reported in [7, 13]. However, in their work they did not take into account the fact that different ads are targeted at different users. Nor did they consider that users differ in their tendency to provide feedback. In other words, users may dislike ads but not indicate this through a feedback option whereas others may always give feedback, however minor the complaint. These potential sources of bias imply that reducing feedback may not equate to improving ad quality for all users.

Any ranking model that utilises such an explicit signal needs to take care when penalizing ads based on a vocal minority. This raises the following two challenges for ad feedback to be used as a quality signal in online advertising applications:

(1) Not all users provide feedback. Can the feedback given be considered reflective of the entire population?
(2) Given the penalization nature of feedback signals in ranking there is naturally a cost attributed in the form of lost short term revenue. Given that publishers will be unwilling to sacrifice such revenue loses, is there a way of finding a tunable compromise between this cost and the quality improvements?

3 AD FEEDBACK DATA

In this paper we use data collected from the feedback mechanism of a popular mobile news app from a major online company.[1]

To illustrate the setup, consider the three anecdotal ads and corresponding feedbacks as shown in Figure 1. Here, users can choose to hide ads they are negatively impacted by simply clicking on the small 'x' appearing in the top right-hand corner of the ad. In the feedback population column we illustrate the types of users who saw a given ad and colour them red when they gave feedback versus black for those who did not. Since the top two ads (a and b) are hidden by more users per impression, they are indicative of a poorer ad quality than the bottom ad (c). However, say for argument's sake that teenage males overall provide more feedback than the rest of the population. For ad (b), since this feedback is attributed only to teenage males, the quality of ad (b) should be higher than ad (a) as the negative feedback attributed to (a) is from more diverse segments of the population.

We extracted a random sample of ad feedback data collected over a two week period from from the US version of the news mobile app. Our sample contains around 40 million distinct users and 200,000 distinct ads. Since feedback rate is a business sensitive metric, we characterize the ad feedback rate relative to ad CTR. The hide rate is 84.0% smaller than CTR, where hide rate is defined as follows:

$$HR = \frac{\#(\text{hides})}{\#(\text{impressions})}$$

where $\#(hides)$ is the number of times users clicked to hide an ad and $\#(impressions)$ is the number of ad impressions.

One may argue that CTR is already an indicator of ad quality. However, feedback is generally a signal of bad quality while the absence of a click does not necessarily indicate a low quality ad. In addition, high CTR may not necessarily mean high quality. As discussed in [30], many ads labelled as offensive could be seen as "provocative", hence attracting clicks. Through hides the user provides an explicit negative signal. A first contribution of this work is to provide insight into whether this signal is representative across users (whether only particular users provide hides), ads (whether only bad ads receive hides) or a combination of these.

The variables (features) and their levels (feature values) used to characterize users are described in Table 1. We distinguish between

[1]The news app was Yahoo news. However the work presented here is relevant to any online sites where ads are served, and an ad feedback mechanism is in place.

Variable	Levels	Dim
	User demographics level partition	
age	below 24; 25-29; 30-34; 35-44; 45-54; 55-59; 60-64; 65+; unknown	9
gender	male; female; unknown	3
location	50 US states	50
interests	business; telecommunications; health; travel; automotive; entertainment; sports; finance; technology; retail; unknown	11
	User behaviour level partition	
impressions	indication of exposure to ads	25
ad clicks	indication of interaction with ads	25
article clicks	indication of interaction with articles	25

Table 1: Variables and levels for each variable used to partition users. The level *unknown* is used when missing data is present.

user demographics and user behaviour based variables. User demographic variables age and gender, are based on users' self declared information (when signing up to the news app). User interests are inferred based on articles clicked by users in the last month using a proprietary in house algorithm. Location is based on the most recent IP address observed for a user.

We categorize user behaviour in terms of the number of ad impressions, ad clicks, and news item (article) clicks. Impressions are a proxy for engagement with the news app, whilst ad clicks and article clicks are proxies for engagement with specific ads or articles, e.g., a user may skim headlines but never read any articles. We bin the counts for each variable on a logarithmic scale into 25 categories. Regarding the features to characterize ads, we were inspired by [30] and selected three types of ad features specific to the ad copy.

- *Text-based* features are derived from the title and description of an ad and include: (a) *spam*, the extent in which the text has spam keywords; (b) *readability*, the readability level according to the Gunning fog index (primary school, high school, etc); (c) *adult*: whether the product advertised is adult-related e.g. dating sites.
- *Image-based* features are derived from any image used in an ad and include: (d) *image text detector*: whether the ad image contains text; (e) *flesh*: likelihood that an image contains adult content (e.g. too much skin).
- *Advertiser* features are derived from the advertiser that placed the ad and include: (f) *brand*: the pagerank score of the top level domain of the ad landing page, reflecting its popularity as a brand.

4 AN ANALYSIS OF AD FEEDBACK

We start by exploring the relationship between user variables and hide rate. To understand which variables may be used to characterize users that hide ads, we study the difference of ad hide behaviour for each user variable in turn. Within each variable e.g. user interests, we define a *cohort* per level, e.g., Travel, aggregate the feedback across each cohort and compare the cohorts. If users within a cohort are targeted with particularly low quality ads compared to other cohorts then we would expect the hide rate to be higher than average for that level of a user variable.

We visualize how the hide rate of various user cohorts $HR(u)$ differ relative to the mean ad hide rate of all user cohorts U. This is defined as:

$$HR_{var}(u) = \frac{HR(u) - mean(HR(U))}{mean(HR(U))}$$

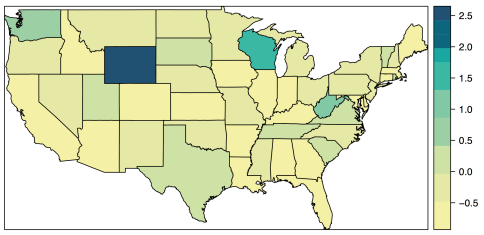
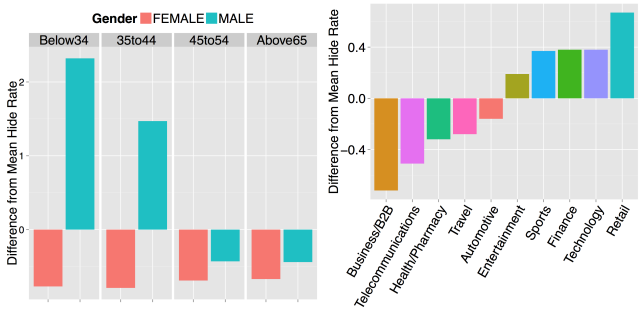


Figure 2: Ad hides across US states



(a) Ad hides across age and gender **(b) Ad hides and user interests**

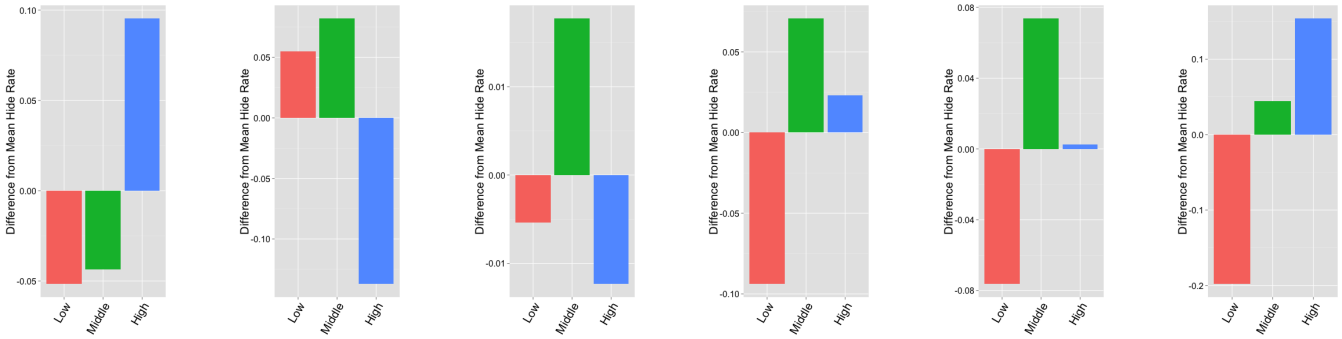
Figure 3: Relation between hides and user variables, i.e., age by gender, and interests.

By visualizing the difference from the mean across levels of each variable we gain insight in whether such a variable may be involved in targeting, e.g., users interested in technology may be more likely to provide ad feedback than users interested in travel, or represent a group of users sensitive to ads.

User Demographics. We first plot the relative difference from the mean HR_{var} according to various US states in Figure 2. We observe that users from different states exhibit differences in hiding ads. This is likely due to targeting as location is a popular targeting criterion. However, demographic distributions differ per state and a generally younger or older state level population may affect hide rate as well.

We next plot HR_{var} across different user age and gender groups in Figure 3a. We observe that female users are less likely to hide ads than male users. With respect to age, young male users tend to hide ads more often, whereas there are no differences across the age groups for female users. Finally, young male users and young female users below the age of 44 are, respectively, most and least likely to hide ads.

User Interests. We plot the HR_{var} across different user interest groups in Figure 3b. We find that users interested in "Retail" and "Technology" are most likely to result in hidden ads. In contrast, users interested in "Business/B2B" and "Telecommunication" are the least likely to provide ad feedback. As with the demographics variables, interest variables are popular targeting criteria and advertisers may target based on one or all of these. We find that feedback variations across the levels of these variables make them suitable candidates to identify bias due to targeting.



(a) Text Spamness (b) Text Readability (c) Text Adultness (d) Image Text (e) Image Flesh (f) Advertiser Brand

Figure 4: Understanding the relationship between ad features and hide rate.

Quantile	Ad Impressions	Ad Clicks	Item Clicks
Q1	-0.38	-0.73	-0.85
Q2	1.89	0.85	1.14
Q3	-0.65	-0.96	0.26
Q4	-0.87	0.84	-0.55

Table 2: User engagement levels and ad feedback

User Engagement Patterns. We are also interested in whether users with different engagement levels with respect to the host content platform respond differently to the ads shown in terms of feedback. We categorize users into four quantiles according to a given engagement metric, where "Q4" represents the highest quantile (the most engaged users) and "Q1" denotes the lowest quantile (the least engaged users).

We use three metrics of user engagement: users' ad impressions, ad clicks and content item (e.g., article) clicks. The first metric is a proxy of how engaged a user is with the news app, as the more engaged, the more the user will have seen ads. The second metric characterizes the extent to which a user engages with an ad, i.e. clicking on it. The final metric is the extent to which the user engages with the content items, i.e. reading news articles.

Table 2 shows how each user behavior metric, at various engagement levels, affects ad hide rates compared to the mean across all users. We observe first that users who are the least engaged with the service and with the ads, are less likely to hide ads (quantile "Q1"). We also see that the second level of engaged users, with respect to all metrics, are those that are more likely to provide ad feedback, i.e. they are more likely to hide ads ("Q2"). Finally, users that are highly engaged with the service tend to hide less ads, but those that interact with ads do provide more ad feedback ("Q4").

The above behavioural variables are not available as targeting criteria. Nonetheless we still observe considerable variations across the levels. For example, those users that sometimes saw ads (Q2 of *ad impressions* in Table 2) are almost as twice as more likely to hide the ads. One explanation is the presence of a response bias for users with different engagement levels, e.g., a user highly engaged with a content platform may be less sensitive to ads shown there than less frequent users affecting their propensity to provide feedback.

Ad quality. High and low values of ad features aimed to characterize the quality of ads should be associated with different levels of ad feedback in order to be useful. Here, we explore the relationship between different levels of ad features and hide rate. We derive ad features from the ad copy (ad text, image and advertiser) and investigate how variants of each affect hide rate. Similarly to our user cohort analysis, we study the difference in hide rates with respect to each ad feature relative to the mean hide rate.

We use the ad features described in Section 3. For each ad feature, we separate all the ads into three equally sized bins ("Low", "Middle" and "High") and visualize how the ads in each bin differ from the mean hide rate of all ads. The results are presented in Figure 4. For text-based features, the more likely an ad has its text identified as *"spam"*, the more likely the ad is hidden (Figure 4a). By contrast, the *easier* the ad is to be *read* (e.g. written at a level associated with primary school students), the more likely it is being hidden (Figure 4b). Finally, ads with the most and least *Text adultness* are less likely to be hidden (Figure 4c), while ads with a moderate level of adult content are most often hidden. In terms of image-based features, the more likely an ad contains *"text"* (Figure 4d) or *"flesh"* (Figure 4e) in its image, the more likely it will be hidden.

Summary. We find that both the user variables and ad features are associated with different levels of ad feedback. Apart from the ad quality, some types of users that provide feedback may be more sensitive to ads and have a higher tendency to provide feedback. In addition, since ads are targeted, the ad feedback received maybe from a group of users unrepresentative of the general population. In the next section, we model this bias considering both users' characteristics (those popular ad targeting criteria, such as age and interests) and their behaviour (e.g., their engagement levels).

5 MODELLING BIAS

The previous analysis provides insights into the distribution of ad feedback across individual user and ad characteristics as well as the presence of bias due to user ad sensitivity and targeting in the ad feedback data. An ad quality model based on such biased data will consistently over- (or under)-estimate the quality of ads. To account for such bias we first, in this section, develop a model able to determine the proportion of bias present in the feedback on ads. Then, in Section 6, we use this model to develop an ad quality model based on ad feedback data corrected for bias.

5.1 Approach

Ideally one would determine the quality of ads based on ad feedback through ad impressions randomized across users. After sufficient impressions, an ad feedback rate would be an unbiased estimate of its true quality. However, since ads may be targeted at users based on specific characteristics or shown to users more prone to

give feedback such data is not available. Hence, we have a sample selection bias problem where ads are shown to a selected sample that nay not be representative of the population.

To correct for this selection bias we do not aim to develop the best *predictive* model for ad feedback, nor do we need to experiment with various features and machine learning models. Although such a model can accurately predict users' feedback it implicitly models the bias present in the data. Instead we take an econometric approach [27] in which we develop a *descriptive* model [25] of user feedback behaviour that only includes variables able to explain the sources of selection bias, i.e., ad sensitivity and targeting. We *will* use this model to account for bias in an *ad quality model* capturing the relation between ad characteristics and ad feedback in Section 6.

To correct for selection bias we use the following procedure. First, consider a simplistic logistic regression based ad-user model with one ad feature a, one user selection bias feature u, associated weights w_a, w_u respectively, the intercept term w_0, the logit link function $f(.)$, and the model error ϵ:

$$f(\hat{p}) = w_0 + w_a \cdot a + w_u \cdot u + \epsilon$$

Further, we consider an ad only model using only one ad feature:

$$f(\hat{p}) = \hat{w}_0 + \hat{w}_a \cdot a + \epsilon$$

If both models are fit to the feedback data with the selection bias present then the bias in the coefficient of the ad model can be characterized as follows:

$$\hat{w}_a = w_a + \rho w_u,$$

where ρ is the correlation between ad feature a and user selection bias feature u. That is, the bias in the ad only model is the true user bias proportional to the correlation between the user and the ad feature. In contrast, in the ad-user model the user selection bias is modelled explicitly and the true effect of the ad feature is captured by w_a, whereas the user selection bias is captured by w_u. Using the w_a and intercept terms from the explicitly biased model, we obtain:

$$f(p) = w_0 + w_a \cdot a + \epsilon$$

This is the feedback rate predicted purely based on the ad characteristics and with weights that are no longer conflated with biased user feedback. Next we develop the user selection bias model and investigate its properties.

5.2 The user selection bias model

We start with developing a model to identify user selection bias term $(w_u \cdot u)$ in the previous equation and then analyze how combinations of characteristics relate to the presence of selection bias in ad feedback. We consider the user characteristics from Table 1 for inclusion in our models and employ a forward model selection based strategy to determine the composition of the final user bias model. As criteria for model fit we use the deviance statistic [16]:

$$D = 2 \sum y_a \log \frac{y_a}{\hat{y}_a} + (N_a - y_a) \log \frac{N_a - y_a}{N_a - \hat{y}_a}$$

where y_a is the observed number of hides and $\hat{y}_a$ is the predicted number of hides for group a.

Table 3 shows the deviance statistics for the models of interest ranging from the null to the best fitting model. The first column lists the model name, the second uses notation from [9] to specify the systemic structure of the model, the third column lists the deviance statistic, the fourth column the difference between the null model and the current model, and finally the fifth column lists the number of parameters used in the model.

model ($\mathcal{M}$)	notation	deviance	$\delta_{\emptyset-\mathcal{M}}\%$	param
Null	$\emptyset$	1438444	-	1
age	a	978274	32	10
gender	g	1413902	2	4
state	s	1169876	19	51
interests	i	788621	45	12
adclk	ac	992030	31	26
itmclk	ic	959711	33	26
adimp	ai	1184212	18	26
additive	a+g+s+i+ac+ic+ai	323621	78	143
interactions	a+s+i+ac+ic+s:i	304426	79	607

Table 3: Deviance statistics for models of interest ranging from the null to the best fitting model for ad hide data of a news app.

We observe that the age variable reduces deviance by 32% (column $\delta_{\emptyset-\mathcal{M}}$) using ten parameters. By contrast, gender reduces deviance by 2% using four parameters and the state variable reduces deviance by 19% using fifty-one parameters. These suggest that different age levels are more indicative of different hide rates then gender or state. However, we find that the interest variable has the best model fit and reduces deviance by 45% using twelve parameters suggesting that interest variables are a popular targeting criteria.

The behavior based variables ad click (adclk) and item click (itmclk) are similar in reducing deviance by 31% and 33% using twenty-six parameters, whereas ad impressions (adimp) reduces deviance by 18%. Moving to the additive model, i.e., combining all variables but not their interactions, reduces deviance by 78% on 143 parameters. However, the best fitting model including two-way interactions achieves a 79% reduction in deviance on 607 parameters. This model does not include gender and ad impressions, but does include an interaction term for state and interests.

Each of the reductions in deviance by adding variables to the models leads to significant improvements as determined using a χ^2 goodness of fit test - i.e., the variables included in the model are able to explain significant structure in the data. There is a meaningful amount of feedback that can be explained by targeting criteria, specifically, age as well as combinations of different levels of state and interest variables regardless of the ad characteristics. These suggest that there is a selection bias due to targeting present in our feedback data. We further observe that click behaviour variables are able to explain additional feedback, suggesting the presence of selection bias due to user ad sensitivity. Next, we describe the effect of specific values of these variables on selection bias.

5.3 User selection bias model inspection

The user selection bias model includes variables that can explain two sources of selection bias, i.e., targeting and ad sensitivity. We now inspect the model to understand how individual variable level weights positively or negatively affect the selection bias present in ad feedback - the *net effects*. We discuss the effects in terms of the odds ratio (odds for short) as it facilitates interpretation of the effect of each variable level on ad hides. We derive the odds ratio from a net effect β using:

$$\frac{p}{1-p} = e^{\beta}$$

When relevant, to determine the probability of hiding ads, we use the inverse logit:

$$p = \frac{e^{\beta}}{1 + e^{\beta}}$$

Table 4 shows the effects of the levels of the variables as produced by the best fitting model. The first column lists the *variables* included

variable	level	net effect	variable	level	net effect
	Intercept	-21.1	adclk	24	-15.7
state	Maine	-0.782▲		11	2.01
	Georgia	1.249▲	itmclk	24	-0.553
interest	Entertain.	-0.104△		1	-1.39
	Retail	0.072	state:interest	OR:Auto	0.898▲
	Finance	1.35 ▲		MT:Auto	0.992▲
	Auto	-0.206		NM:Telecom	1.124△
	Telecom	-0.759▲		SD:Retail	1.637▲
				WY:Auto	1.760▲

Table 4: Net effects of variables in the best model. Significant effects are denoted by △(▲) at the $\alpha < 0.05$ ($\alpha < 0.01$) level.

variable	level	effect	variable	level	effect
	intercept	-9.203 ▲	user type:adult score	L1:L10	-1.144 ▲
user type (u)	L1	3.012 ▲		L1:L8	-0.632 ▲
readability (r)	L1	0.082 △		L1:L6	-0.586 ▲
	L4	0.204 ▲		L1:L2	-0.217 ▲
adult score (a)	L9	-0.187		L1:L1	-0.158 ▲
	L8	0.967 ▲	user type:pagerank	L1:L79	-2.498 ▲
spam score (s)	L7	-0.194 ▲		L1:L89	-0.401 ▲
	L9	2.058 ▲	pagerank:adult score	L79:L8	-3.478 ▲
pagerank (p)	L89	-0.993 ▲	pagerank:spam score	L79:L7	0.7758 ▲
	L79	3.270 ▲			

Table 5: Selection of the effects of the ad and user variables in the best fitting model for the ad hide data. In formula notation the model is specified as follows: u+p+a+r+s+u:p+u:a+p:a+p:r+p:s+a:r+a:s. Significant effects are denoted by △(▲) at the $\alpha < 0.05$ ($\alpha < 0.01$) level.

in the model, the second column contains the *levels* of each variable, and the third column presents the *effects* of each level.

The intercept is negative and the odds ratio of hiding ads when controlling for all other factors is small ($6.86 \cdot 10^{-10}$). It reflects the hide rate of a user with no targeting characteristics and having the lowest engagement level in terms of ad click, item click, and ad impression behaviour. We will refer to this user as our reference user on the news app. This user's hide rate is orders of magnitude smaller than the general feedback rate (see Section 3) and implies that the reference user is unlikely to hide ads. The characteristics of our reference user are based on the levels that are used as reference cell for each variable, i.e., a user with no targeting characteristics and having the lowest engagement level in terms of ad click, item click, and ad impression behaviour. We discuss the variables related to targeting and user ad sensitivity each in turn.

Targeting effects. Table 4 shows the two states with the largest negative and positive effects. We observe that the net effect of changing our reference user state to Maine reduces the odds of observing a hide by 64% ($1 - e^{-0.782}$), whereas changing to Georgia increases the odds by 249% ($e^{1.249}$). This is in line with our observations in Section 4, Figure 2. In terms of the interest levels we find that an interest in Sports and Finance increases the odds of observing a hide by 150% and 284%, respectively. In contrast, interest in the Telecommunications category reduces the ad hide odds by 63%. These observations are again in line with our findings in Section 4, Figure 3b. We further discover that there are interaction effects; Table 4 shows the five levels with the highest positive net effect. Three of the five highest effects are combined forces between an interest in the Automotive category and some of the states, i.e., increasing the ad hide odds by 145% in Oregon, 151% in Montana, and 481% in Wyoming, respectively.

Behavioural effects. The behavioural variables capture the extent in which users interact with ads and items (e.g., articles). The interaction data for each user is binned on a logarithmic scale in 25 categories. Table 4 shows the 2 levels with the highest negative and positive effects. Moderate ad clickers are more likely to hide an ad than users with no ad click activity or very high ad click activity.

Further, there is an inverse linear relation between the number of ad hides and the 25 log scale levels of the itmclk variable (Pearson $r = -0.99$, $p < 0.001$). This implies that, unlike for ad clicks, users clicking on more articles are less likely to hide ads, whereas users clicking on few articles are more likely to hide ads. A user having an itmclk level of 24 reduces the odds of an ad hide by 42% and a user with an itmclk level of 1 increases the ad hide odds by 300%.

Discussion. Inspection of the user selection bias model shows that not all users hide ads equally as user characteristics alone are sufficient to explain a significant part of the feedback observed in our data without considering features indicating the quality of ads. Ad feedback varies depending on particular user targeting characteristics, which demonstrates that ads from certain ad campaigns with targeted, for example, towards users in specific locations with particular interests received high levels of feedback. Using feedback only from such particular groups of users is not appropriate to train a general ad quality model that determines whether ads are of high or low quality for the general population. Feedback also varies depending on the engagement level of users of the news app; for example, users with high engagement levels with articles were less likely to provide feedback than users with low engagement levels. It seems reasonable to trust the feedback received from users over various levels of engagement more than feedback only provided by a group of users with low engagement levels.

Next, we incorporate the user selection bias model into an ad quality model that uses the user selection bias as well as ad features to determine the hide rate of ads.

6 CORRECTING BIAS IN AD FEEDBACK

To correct for selection bias in the feedback used by our ad quality model we explicitly model the user selection bias in addition to the ad features as follows:

$$f(\hat{p}) = w_0 + w_a \cdot a + I(w_u \cdot u) + \epsilon$$

We use an indicator function to binarize the user selection bias term based on a threshold parameter $I : f(.) > t$. This parameter indicates whether feedback received on an ad is likely due to selection bias.

We consider four ad characteristics as variables in our model, namely, pagerank of domain (pr), adult score (a), readability description (r), and spam score (s). Given each ad feature distribution we separate them into equally sized bins representing the different levels. We use 100 levels for pagerank of domain, i.e., L1 to L100, 5 levels for the readability score and 10 levels for the remaining features. We do not include the ad image features as their distributions are more skewed, i.e. only a small number of ads contain text or flesh. Further, we set the threshold for the user selection bias model to 0.5 to indicate when part of the feedback received is likely to be due to selection bias and encode this as level L1 when $f(w_u \cdot u) > 0.5$ and L0 otherwise. Given the above variables we again use forward stepwise model selection based on AIC and allow inclusion of the main effects and all pair-wise interactions in the candidate models.

The effects of the variable levels of the resulting model are presented in Table 5. The model includes the main effects of user bias (u), page rank of domain (p), adult score (a), readability (r), and spam score (s). In terms of interaction effects, the model includes

all pairwise interactions between page rank and the other variables (u:p, p:r, p:s, p:a,) and all interactions between adult score and the other variables (u:a, a:r, a:s). We observe that the hide rate derived from the intercept is 0.01%, which is several orders of magnitude higher than the overall feedback rate, see Section 3. This indicates that ads with low pagerank and readability levels as well as low adult and spam levels are considered of low quality by users when correcting for selection bias. Adding the effect of the user selection bias variable (3.012) results in a further 20 fold increase in the odds of observing a hide, suggesting that users specifically targeted by ads with such features and users with a high disposition to give feedback also consider such ads of low quality.

With respect to the other main effects, we find that the effects have mixed directions across the levels of each variable. The readability levels with significant effects all result in an increase in the odds of observing a hide; however, lower levels do less so than higher levels. The highest and lowest significant effects for the adult score are both for high levels of adultness. A similar observation holds for the spam score and pagerank. These observations reflect the presence of interaction effects between levels of ad features, i.e., single features do not characterize the quality of an ad.

In terms of interaction effects we observe that ads with some features are less likely to receive biased feedback than others. For example, an ad with adult level L8 increases the odds of receiving a hide by 163%; however, if the ad is impressed to a user of level L1 then the odds are reduced. Similarly, an ad with a high pagerank level, i.e., L79 increases the odd of receiving a hide by a factor of 36; but impressed to a user of level L1, those odds reduce by 92%. Ads with such features are likely to receive feedback from a general population, but unlikely to receive feedback from a specific segment of users. This may be indicative of ads with products or content prohibited in certain countries, demographics, etc.

The effects for the ad features in the bias corrected ad model determine the inherent quality of an ad as determined by the hide rate estimated for the general population and not a specific segment of users. By adding all effects based only on the characteristics of each ad — excluding the user selection bias factors and interactions — we obtain the unbiased hide rate of an ad that controls for the type of users that ad has been shown, i.e., $f(p) = w_0 + w_a \cdot a + \epsilon$, from Section 5. In the next section we show the utility of correcting for user selection bias in ad feedback in ad ranking.

7 APPLYING FEEDBACK IN AD RANKING

We first introduce the general ad ranking approach for online ad auctions as well as how ad quality scores are generally incorporated in such rankings. Then we describe our method that allows the use of ad quality scores in ad ranking under a fixed revenue constraint. With these in place, we compare the value that biased and bias corrected feedback estimates provide in terms of discounting or filtering bad quality ads with particular attention to how this impacts revenue.

7.1 Ad auctions and feedback data

Online advertisement impressions are sold using real-time auctions. In such auctions each time a user loads a webpage one or more ad slots become available for sale. Advertisers state their bids for an impression of one of these slots and ads are ranked according to a function of their bids. The ads with the highest ranks are displayed and if a user clicks on the ad the advertiser is charged an amount based on the bid of the ad directly below it in the ranking [3]. As it directly influences prices charged to advertisers and therefore the revenue of the host site, the ranking function is central to the auction.

Ads are generally ranked by expected cost per impression (eCPI), which is determined based on the probability that the ad (a) is clicked given a user impression (u):

$$\text{eCPI} = \text{bid}_a \cdot P(C_a = 1|U = u)$$

The immediate revenue is maximized by selecting the ad a from a set of ads A eligible for an impression during a real time auction with highest eCPI:

$$\arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|U = u) \tag{1}$$

where the probability of a click on ad a given user u may be estimated based on historical click data and preferences of similar users [1].

Some Internet companies have started to incorporate quality scores into their ad ranking mechanisms. We define the probability that an ad will provide a good quality experience as: $P(Q_a = 1|U = u) = 1 - p$, where p is an estimate that the user u will hide the ad. One common method to incorporate quality into selecting which ad to show to users is to discount the expected CPI by the probability that the ad will deliver a quality experience [20]. In the case of ad feedback we can formulate discounted eCPI as:

$$eCPI_q = \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u)$$

and select the ad that maximizes revenue:

$$\arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u) \tag{2}$$

Results. We use two weeks of ad interaction and revenue data from the logs of the Internet company for which we study the ad feedback data to demonstrate the impact of incorporating feedback data (quality) in ad ranking. We compare the revenue delivered and negative feedback (hides) prevented over a two week period for three ad feedback based quality ranking methods (implemented using equation 2) with the performance of the current eCPI based model (i.e., equation 1) running in production of the Internet Company.

Figure 5 shows the % change in quality improvement (hides prevented in grey) and revenue delivered (in red) for three ways of estimating $P(Qa = 1|U = u)$: (i) *oracle*, which uses the emperical estimate based on the logs in hindsight, (ii) *biased*, which uses estimates based on the biased model $f(\hat{p})$ from Section 5; and (iii) *unbiased*, which uses estimates based on corrected model from Section 6. We observe that the general approach to incorporating quality scores into the ad ranking function does not deliver a favorable trade-off in terms of quality improvement and revenue investment. The quality ranking approach using oracle estimates of the quality score (hide rate) results in a 28.8% reduction in revenue, while reducing the number of hides by 16.7%. The biased and unbiased estimate based approaches achieve similar performance and are worse reaching a 36% and 37% revenue investment respectively for a 18% reduction in hides.

Next, we explore alternative ad ranking approaches that take revenue into account.

7.2 The revenue-feedback trade-off

When a publisher maximizes expected CPI that is discounted by the quality of ads it no longer maximizes revenue, unless the quality is proportional to eCPI or uniformly distributed. Specifically, given a set of ad candidates $A = \{a_1, ..., a_n\}$ for an impression we define the maximal expected revenue as:

$$\text{eCPI}_{max}(A) = \text{bid}_{a_i} \cdot P(C_{a_i} = 1|U = u)$$

where $a_i = \arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|U = u)$.
Further, we define the quality based expected revenue as:

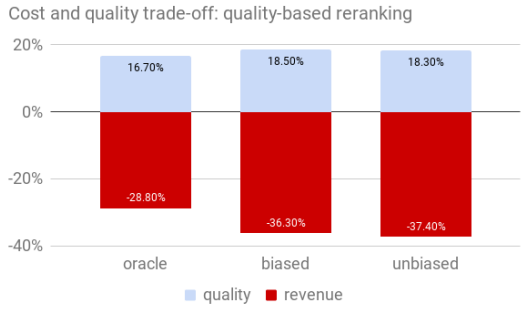


Figure 5: The % change in hides prevented (grey) and revenue delivered (in red) for three ways of estimating $P(Qa = 1|U = u)$: (i) *oracle*, using emperical log based estimates, (ii) *biased*, which uses predicted estimates based on the biased model $f(\hat{p})$ from Section 5; and (iii) *unbiased*, which uses predicted estimates based on corrected model from Section 6.

$$\text{eCPI}_q(A) = \text{bid}_{a_j} \cdot P(C_{a_j} = 1|U = u)$$

where

$$a_j = \arg\max_{a \in A} \ \text{bid}_a \cdot P(C_a = 1|Q_a = 1) \cdot P(Q_a = 1|U = u)$$

Then the loss in eCPI is equal to

$$\text{eCPI}_{loss}(A) = \text{eCPI}_{max}(A) - \text{eCPI}_q(A)$$

The gain of showing quality ads depends on an increase in impressions and clicks compared to not showing quality ads over a certain period of time. However, any increase in impressions and clicks, and therefore any benefit of showing quality ads, can only be observed in hindsight. We define the set of impressions that would have been observed while maximising for expected CPI during window (t, t_k) as $I_{(t,t_k)}$ and the impressions potentially observed while discounting for quality as $I_{q(t,t_k)}$, where we assume $I_{(t,t_k)} \subseteq I_{q(t,t_k)}$, i.e., quality ads indeed lead to more engaged users. Publishers then wish to find a time t_k where:

$$\sum_{j \in I_{q(t,t_k)}} \text{eCPI}_q(A_j) - \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i) \geq 0$$

By splitting the impressions in those that would have been observed under the regular ad serving model $I_{(t,t_k)}$ and potential new impressions $I_q \setminus I$ we obtain:

$$\sum_{j \in I_q \setminus I} \text{eCPI}_q(A_j) + \sum_{i \in I_{(t,t_k)}} \text{eCPI}_q(A_i) - \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i) \geq 0$$

Rewriting we find that publishers face the dilemma of whether to risk a loss in revenue

$$\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i)$$

in order to find if there is a time t_k where:

$$\sum_{j \in I_q \setminus I} \text{eCPI}_q(A_j) \geq \sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i) \tag{3}$$

while having no control over the revenue loss

$$\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{loss}(A_i).$$

7.3 Bounding short-term revenue loss

Whether the left hand side of Equation 3 will become greater than the right hand side depends on the relation between ad quality and user engagement as well as the magnitude of the bids associated with quality ads. To allow publishers to investigate whether quality ads could yield more revenue in a particular market place while being in control of the amount of revenue lost, we propose the following:
1) Select all the ads A that are eligible to be shown to the user;
2) Compute the maximum expected revenue given A: $\text{eCPI}_{max}(A)$;
3) Find the ad with maximum user experience such that the loss in revenue is smaller than a threshold τ:

$$\begin{aligned} a_c &= \arg\max_{a \in A} \ P(Q_a = 1|U = u) \\ \text{s.t.} &\frac{(\text{eCPI}_{max}(A) - \text{bid}_a \cdot P(C_a = 1|U = u))}{\text{eCPI}_{max}(A)} \leq \tau \end{aligned} \tag{4}$$

Given Equation 4 we define *the controlled quality discounted eCPI*

$$\text{eCPI}_c(A) = \text{bid}_{a_c} \cdot P(C_{a_c} = 1|U = u)$$

Substituting regular quality discounted eCPI for controlled quality discounted eCPI and normalizing by the total revenue for eCPI_{max} in Equation 3 we obtain:

$$\frac{\sum_{j \in I_q \setminus I} \text{eCPI}_c(A_j)}{\sum_{i \in I_{(t,t_k)}} \text{eCPI}_{max}(A_i)} \geq \tau \tag{5}$$

By controlling the maximum loss in revenue on a per impression basis we limit the potential revenue loss to a fixed percentage. This allows publishers to predetermine the amount of risk they wish to take in exploring whether introducing quality ads in a market place will yield an increase in revenue over a standard serving model.

7.4 Impact of ad feedback filtering on revenue

As demonstrated in Figure 5, it may be costly to evaluate the revenue impact of showing users quality ads as it is unknown when (or if) the initial revenue investment (showing high quality but less monetizable ads) is offset by an increase in user engagement [15]. To investigate whether unbiased ad feedback estimates provide additional value over biased estimates we evaluate their effectiveness in reducing hide rate at different levels of short-term revenue investment. We again extract two weeks of ad interaction and revenue data from the logs of the Internet company. For each first slot auction we find the top 10 ads that participated based on $eCPI_{max}$.

We follow a simulation based approach based on [2] and simulate for each impression the ad that would have been shown using either a biased estimate (i.e., the full model from Section 6) or a bias corrected estimate (i.e., only the ad coefficients from Section 6), as well as the true hide rate for each auction using Equation 4.

Figure 6 shows the proportional revenue loss threshold, i.e., the percentage short term revenue investment on the x-axis (defined as τ in Section 7.3) and the reduction in ad feedback for each method at each threshold on the y-axis. For each threshold $\tau \in \{1, 5, 10\}$ percent short term revenue investment, we determine the ads shown in the simulated auctions for each method and find the reduction in hide rate compared to the original auctions.

We observe that at 1% short term revenue investment there is little difference between the biased and unbiased curves, implying that the initial ads that are filtered are found to be bad by all users. This is an important finding, as it shows that for the worst of the bad ads, ad feedback data is not biased and can be used as is to remove bad ads. Furthermore, the ads (and their hides) that are removed by the bias corrected estimates are received from all users, suggesting they are considered of low quality by the general population. While

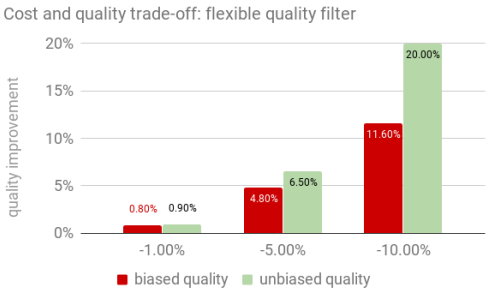


Figure 6: Revenue investment-hide reduction trade-off as rate of total when filtering ads using eqn. 4 based on biased (red), and unbiased (green) estimates of hide rate.

ads removed based on the original quality estimates may have been received from only a particular sub-population.

We observe the true benefit of unbiased ad feedback estimates at a 10% revenue investment as at that point a 20% reduction in hides can be achieved compared to about a 12% reduction using biased estimates. Moreover, at 10% the short term revenue investment is far lower than the 28% to 37% cost when using the traditional quality based ranking approach while achieving higher quality improvements, cf. Figure 5. Based on business criteria other trade-offs may be selected; however, our results suggest that there is value in using unbiased estimates in that they provide the best return on investment.

8 CONCLUSIONS AND FUTURE WORK

Publishers relying on advertising for revenue are aware that optimizing for CTR is not sufficient for long-term engagement. They must also consider the quality of the ads shown to their users. To this end, some provide mechanisms to gather explicit feedback about the ads. This work is the first providing insights into the sources of bias present in ad feedback data, i.e. hide rate, in estimating ad quality.

Using hide rate, we focused on understanding and modeling the characteristics of users who do not want to see ads. An initial analysis of the ad feedback data shows that some types of users may be exposed to lower quality ads than others and as such result in higher feedback rates. More precisely, we found evidence for the presence of two sources of selection bias, i.e., targeting and response bias. To correct for such bias we develop a so-called user selection bias model that allowed us to quantify the bias present in the feedback on ads, considering both users' characteristics (ad targeting criteria, such as age and interests) and their behaviour (engagement levels).

We then incorporate the user selection bias model into an ad quality model, which results in a bias corrected ad model that provides a true estimate of the inherent quality of the ads. The corrected estimate is the unbiased hide rate of an ad that controls for the type of users that ad has been shown. Further, we introduced a method to control revenue investment when using ad feedback estimates in practice. We found that when comparing the use of biased ad feedback estimates with bias corrected ones in a business scenario that unbiased estimate provide a benefit in terms of the quality-revenue trade-off. However, these benefits only materialize at certain levels of revenue investment and with a mechanism that is able to carefully control revenue investment. Our work allows Internet companies to experiment with ad quality at a fixed risk level, lowering the threshold to explore further quality signals as well as ranking techniques to improve users' experience with ads.

We acknowledge that our findings are based on a sample of data from one Internet company, Yahoo news app. However, the size of our sample and the use of features in our analyses are generally available in the advertisement industry, therefore providing ad networks and publishers with the insights necessary to mitigate user bias in ad feedback data.

REFERENCES

[1] M. Aharon, N. Aizenberg, E. Bortnikov, R. Lempel, R. Adadi, T. Benyamini, L. Levin, R. Roth, and O. Serfaty. Off-set: one-pass factorization of feature sets for online recommendation in persistent cold start settings. In *RECSYS'13*, pages 375–378. ACM, 2013.
[2] A. Ashkan and C. L. A. Clarke. Modeling browsing behavior for click analysis in sponsored search. In *CIKM*, 2012.
[3] S. Athey and D. Nekipelov. A structural model of sponsored search advertising auctions. In *Sixth ad auctions workshop*, volume 15, 2010.
[4] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern. Visual appearance of display ads and its effect on click through rate. In *CIKM*, 2012.
[5] N. Barbieri, F. Silvestri, and M. Lalmas. Improving post-click user engagement on native ads via survival analysis. In *WWW'16*, pages 761–770, 2016.
[6] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: Quantifying the impact of landing pages in web advertising. In *CIKM'09*, 2009.
[7] G. Brajnik and S. Gabrielli. A review of online advertising effects on the user experience. *IJHCI*, 26(10):971–997, 2010.
[8] A. Z. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *CIKM'08*, 2008.
[9] J. M. Chambers and T. J. Hastie. *Statistical models in S*. CRC Press, Inc., 1991.
[10] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *SIGKDD'12*, pages 777–785. ACM, 2012.
[11] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using landing pages for sponsored search ad selection. In *WWW'10*, 2010.
[12] H. Cramer. Effects of ad quality & content-relevance on perceived content quality. In *CHI '15*, pages 2231–2234, New York, NY, USA, 2015. ACM.
[13] D. G. Goldstein, R. P. McAfee, and S. Suri. The cost of annoying ads. In *WWW'13*, 2013.
[14] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, R. Baeza-Yates, A. Feng, E. Ordentlich, L. Yang, and G. Owens. Scalable semantic matching of queries to ads in sponsored search advertising. In *SIGIR'16*, pages 375–384, 2016.
[15] H. Hohnhold, D. O'Brien, and D. Tang. Focusing on the long-term: It's good for users and business. In *KDD'15*, pages 1849–1858, 2015.
[16] D. W. Hosmer Jr and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.
[17] S. Ieong, M. Mahdian, and S. Vassilvitskii. Advertising in a stream. In *WWW'14*, 2014.
[18] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov. Categorization of display ads using image and landing page features. In *LDMTA*, 2011.
[19] D. Kempe and B. Lucier. User satisfaction in competitive sponsored search. In *WWW'14*, 2014.
[20] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users.
[21] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users. In *SIGKDD'15*, pages 1929–1938, 2015.
[22] R. J. Oentaryo, E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *WSDM'14*, 2014.
[23] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *WSDM'12*, 2012.
[24] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *KDD'09*, 2009.
[25] G. Shmueli. To explain or to predict? *Statistical science*, pages 289–310, 2010.
[26] E. Sodomka, S. Lahaie, and D. Hillard. A predictive model for advertiser value-per-click in sponsored search. In *WWW'13*, 2013.
[27] F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.
[28] T. Wang, J. Bian, S. Liu, Y. Zhang, and T.-Y. Liu. Psychological advertising: exploring user psychology for click prediction in sponsored search. In *SIGKDD'13*, 2013.
[29] D. Yin, B. Cao, J.-T. Sun, and B. D. D. 0001. Estimating ad group performance in sponsored search. In *WSDM'14*, 2014.
[30] K. Zhou, M. Redi, A. Haines, and M. Lalmas. Predicting pre-click quality for native advertisements. In *WWW'16*, pages 299–310, 2016.