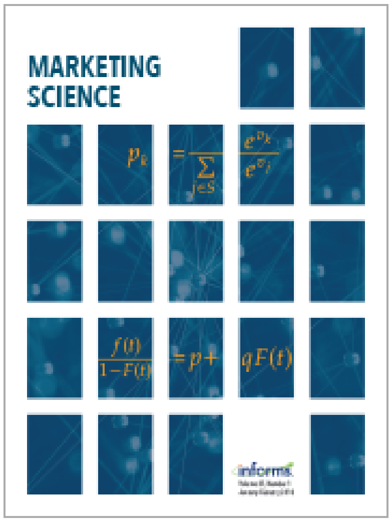


Marketing Science

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

Targeting and Privacy in Mobile Advertising

Omid Rafieian, Hema Yoganarasimhan

MARKETING SCIENCE
Vol. 40, No. 2, March–April 2021, pp. 193–218
ISSN 0732-2399 (print), ISSN 1526-548X (online)
http://pubsonline.informs.org/journal/mksc

Targeting and Privacy in Mobile Advertising

Omid Rafieian,[a] Hema Yoganarasimhan[b]

[a] Cornell Tech and SC Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853; [b] Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195

Contact: or83@cornell.edu, https://orcid.org/0000-0001-8633-2302 (OR); hemay@uw.edu, https://orcid.org/0000-0003-0703-5196 (HY)

Received: April 13, 2018
Revised: April 9, 2019; January 30, 2020; April 8, 2020
Accepted: April 18, 2020
Published Online in Articles in Advance: December 8, 2020

https://doi.org/10.1287/mksc.2020.1235

Abstract. Mobile in-app advertising is now the dominant form of digital advertising. Although these ads have excellent user-tracking properties, they have raised concerns among privacy advocates. This has resulted in an ongoing debate on the value of different types of targeting information, the incentives of ad networks to engage in behavioral targeting, and the role of regulation. To answer these questions, we propose a unified modeling framework that consists of two components—a machine learning framework for targeting and an analytical auction model for examining market outcomes under counterfactual targeting regimes. We apply our framework to large-scale data from the leading in-app ad network of an Asian country. We find that an efficient targeting policy based on our machine learning framework improves the average click-through rate by 66.80% over the current system. These gains mainly stem from behavioral information compared with contextual information. Theoretical and empirical counterfactuals show that although total surplus grows with more granular targeting, the ad network's revenues are nonmonotonic; that is, the most efficient targeting does not maximize ad network revenues. Rather, it is maximized when the ad network does not allow advertisers to engage in behavioral targeting. Our results suggest that ad networks may have economic incentives to preserve users' privacy without external regulation.

History: Puneet Manchanda served as the senior editor and Olivier Toubia served as associate editor for this article.
Supplemental Material: The data and the online appendices are available at https://doi.org/10.1287/mksc.2020.1235.

Keywords: advertising • mobile • machine learning • targeting • behavioral targeting • privacy • auctions • regulation • public policy

1. Introduction

1.1. Mobile Advertising and Targeting

Mobile advertising now constitutes the largest share of total digital ad spend (Enberg 2019). The popularity of mobile advertising stems from an ad format unique to the mobile environment: in-app ads or ads shown inside apps. These ads have excellent user-tracking properties and allow ad networks to stitch together user data across sessions, apps, and advertisers.[1] Thus, one of the main attractions of in-app advertising is its ability to facilitate behavioral targeting (Edwards 2012).

Whereas the advertising industry has lauded the trackability of in-app ads, consumers and privacy advocates have derided them, citing privacy concerns. Advertisers argue that tracking allows consumers to enjoy free apps and content and see relevant ads, whereas users demand higher privacy and limits on behavioral tracking and targeting (Edwards-Levy and Liebelson 2017). Responding to consumer concerns, regulatory bodies have started taking action. For example, the European Union's General Data Protection Regulation requires users to opt into, rather than opt out of, behavioral targeting (Kint 2017).

Even as consumers, businesses, and regulators are trying to find the right balance between consumer protection and business interests, we do not have a good understanding of the key issues at the core of targeting and privacy. For example, to what extent does targeting improve the efficiency of the advertising ecosystem, what is the value of different types of targeting information, and what are the incentives of different players in the advertising industry to engage in user tracking and behavioral targeting? The lack of a cohesive framework to analyze these issues hampers our ability to have an informed discussion and to form policy on them.

1.2. Research Agenda and Challenges

In this paper, we seek to address this gap by providing a unifying framework to answer the following sets of questions related to targeting and privacy in the advertising ecosystem. The first set of questions relates to targeting and efficiency. How can ad networks use the data available to them to develop targeting policies? How can we evaluate the performance of these policies in both factual and counterfactual settings? In particular, what are the gains in click-through rate (CTR) from adopting an efficient (CTR-maximizing) targeting policy?

The second set of questions relates to the value of targeting information. We are particularly interested in the relative value of contextual versus behavioral information. The former captures the context (when and where) of an impression, and the latter summarizes an individual user's past app usage, ad exposure, and ad response. Contextual information is privacy preserving, whereas behavioral information is based on user tracking and therefore impinges on users' privacy.

Third, we are interested in quantifying the revenue–efficiency trade-off and ad network's incentives to enable different forms of targeting. What is the empirical relationship between efficiency and ad network revenues? What is the optimal level of targeting from the perspective of different players in the market? Finally, to what extent are the ad network's and advertisers' incentives aligned?

There are three main challenges that we need to overcome to satisfactorily answer these questions. First, to develop efficient targeting policies, we need to obtain accurate estimates of CTR for all ads that could have been shown in an impression (i.e., counterfactual ads) and not just the ad that was actually shown in that impression. Thus, we need exogenous variation in the ad-allocation mechanism to evaluate counterfactual targeting policies. Second, to quantify the value of different pieces of targeting information, we need a model that can accurately predict whether a targeted impression will lead to a click or not. Models with poor predictive ability will lead to downward bias in the estimates of the value of information. Third, we need an underlying model of strategic interactions that can quantify market outcomes (e.g., ad network and advertiser revenues) under different targeting regimes. Without an economic model that puts some structure on the ad network's and advertisers' utilities, we cannot make any statements on their incentives to target and/or the extent to which these incentives are aligned.

1.3. Our Approach

We present a unified and scalable framework that coherently combines predictive machine learning models with prescriptive economic models to overcome the challenges listed earlier. Our framework consists of two main components. The first, a machine learning framework for targeting, addresses the first two challenges of obtaining counterfactual CTR estimates and achieving high predictive accuracy in this task. The second is an analytical model that incorporates competition and characterizes the ad network's and advertisers' profits under different targeting regimes. This addresses the third challenge of linking targeting regimes to ad network and advertiser revenues.

The main goal of the first component is to estimate the match value between an impression and an ad, where match value can be interpreted as the CTR of an impression–ad combination. Once we have match values for all impression–ad combinations, we can use them to define and evaluate any counterfactual targeting strategy. Match values are thus the key primitives of interest here, and we infer them by combining ideas from causal inference with predictive machine learning models. Our approach consists of three parts: (1) a filtering procedure, (2) a feature-generation framework, and (3) a learning algorithm. The goal of the filtering procedure is to identify the set of ads for which we can generate accurate counterfactual estimates of CTR for each impression. If the platform uses a deterministic ad-allocation mechanism (as is common practice in the industry), then this set is null, by definition. However, in our setting, there is exogenous variation in the ad-allocation process, which gives us a nonempty set of counterfactual ads for each impression. Our filtering procedure determines this set by identifying the ads that have a nonzero propensity of being shown in a given impression. Next, our feature-generation framework relies on a set of functions to generate a large number of features that capture the contextual and behavioral information associated with an impression. Using these functions, we generate a total of 160 features that serve as input variables into a CTR prediction model. Finally, we use XGBoost, proposed by Chen and Guestrin (2016), a fast and scalable version of boosted regression trees, as our learning algorithm.

In the second component, we focus on the ad network's incentives to allow targeting. In an influential paper, Levin and Milgrom (2010) conjecture that whereas high levels of targeting can increase efficiency in the market, it can reduce the ad network's revenue by softening the competition between advertisers. We propose a theoretical framework that allows us to characterize this revenue–efficiency trade-off under counterfactual targeting regimes. To take this framework to data, we need an estimate of each advertiser's valuation for a given impression. This valuation can be decomposed into two sets of primitives: (1) match valuations or CTRs for all impression–ad combinations and (2) advertisers' click valuations for each impression. Although match valuations are already available from the machine learning targeting framework, we need to infer click valuations by inverting advertisers' observed equilibrium bids (Guerre et al. 2000). The product of these two entities gives us each advertiser's value of a given impression, which allows us to quantify the ad network's revenue, advertisers' surplus, and total surplus under different targeting regimes.

We apply our framework to one month of data from the leading mobile ad network from a large Asian country. The scale and scope of the data are large enough to provide realistic substantive and counterfactual results. For our analysis, we sample over 27 million impressions for training, validation, and testing and use another 146 million impressions for feature generation. A notable feature of our setting is that a quasi-proportional auction allocates impressions to ads using a probabilistic rule: an advertiser's probability of winning an impression is proportional to his or her bid. This induces randomization or exogenous variation in ad allocation, which, in turn, allows us to estimate match valuations for counterfactual ad–impression combinations. At the same time, the auction mechanism preserves the strategic linkage between bids and advertisers' click valuations, which allows us to estimate click valuations from the bid data. Our setting thus facilitates the separate identification of both match and click valuations.

1.4. Findings and Contribution

We first discuss the results from the machine learning model for targeting. We present both factual and counterfactual evaluations of our model. In the factual evaluation, we use goodness-of-fit measures to evaluate how well our model can predict the observed outcome. We find that our model predicts the outcome on a hold-out test set with substantial accuracy: It achieves a Relative Information Gain (RIG) of 17.95% over a baseline model that simply predicts the average CTR for all impressions. Next, we find that behavioral information contributes more to the predictive accuracy of the model than contextual information. In the second part of our evaluation, we consider the efficient targeting policy, wherein each impression is allocated to the ad with the highest estimated CTR in that impression. We show that this efficient targeting policy can increase the average CTR in the ad network by 66.80% over the current system.

Next, we link advertisers' targeting strategies to the ad network's incentives and revenues. First, we theoretically prove that in an efficient auction mechanism (e.g., second-price auction), (1) the total surplus in the system monotonically increases as the extent of targeting increases, but (2) the ad network's revenues are not monotonic; revenue may or may not increase with more granular targeting. Thus, we take our theoretical framework to data and perform empirical counterfactuals to compare ad network revenues under different targeting regimes.

In particular, we consider four targeting regimes that relate to our research agenda: full (impression-level targeting), behavioral (user-level targeting), contextual (app-time-level targeting), and no targeting. We find that the ad network's revenue is maximized when it restricts targeting to the contextual level even though doing so lowers total surplus; that is, allowing behavioral targeting thins out the market, which, in turn, reduces ad network revenues. Therefore, the ad network has economic incentives to adopt a privacy-preserving targeting regime, especially if it cannot extract additional surplus from advertisers through other mechanisms. On the advertisers' side, we find that although a majority of them prefer a regime where the ad network allows behavioral targeting, not all do. An important implication of our findings is that it may not be necessary for an external entity such as the European Union or Federal Trade Commission to impose privacy regulations in light of ad networks' economic incentives.

Our paper makes several contributions to the literature. First, from a methodological perspective, we propose a novel machine learning framework for targeting that is compatible with counterfactual analysis in a competitive environment. A key contribution of our targeting framework is in combining existing ideas from causal inference literature with recent machine-learning literature to generate counterfactual estimates of user behavior under alternative targeting regimes. Further, we present an efficient auction framework with targeting that characterizes advertisers' utility function under any targeting regime and provides a direct link to the estimation of market outcome such as efficiency and revenue. Second, from a substantive perspective, we provide a comprehensive comparison of contextual and behavioral targeting, with and without the presence of competition. To our knowledge, this is the first study to compare the role of behavioral and contextual targeting on market outcomes. Third, from a managerial perspective, our results demonstrate a nonmonotonic relationship between targeting granularity and revenue. Although our findings may depend on the context of our study, our framework is generalizable and can be applied to most standard advertising platforms that use deterministic auctions as long as the platform randomizes ad allocation over a small portion of the traffic (which would satisfy the unconfoundedness assumption). Finally, from a policy perspective, we identify the misalignment of the ad network's and advertisers' incentives regarding behavioral and contextual targeting and information disclosure. We expect our findings to be of relevance to policymakers interested in regulating user tracking and behavioral targeting in the advertising space.

The rest of this paper is organized as follows. In Section 2, we discuss the related literature. We introduce the setting and data in Section 3. In Section 4, we present our machine learning framework for targeting, and in Section 5, we present a series of results on efficiency gains from targeting. Next, in Section 6, we develop a theoretical framework for analyzing the revenue–efficiency trade-off and a corresponding empirical analysis of auctions with targeting. In Section 7, we present the results on market outcomes under counterfactual targeting regimes. Finally, in Section 8, we conclude with a discussion on the generalizability of our framework and our main contributions.

2. Related Literature

First, our paper relates to the computer science literature on CTR prediction (McMahan et al. 2013, He et al. 2014, Chapelle et al. 2015). These papers make prescriptive suggestions on feature generation, model selection, learning rates, and scalability. Our work differs from these papers in two main ways. First, we develop a filtering procedure that allows us to obtain accurate CTR estimates for both the ad shown during an impression as well as counterfactual ads not shown in the impression. Thus, unlike the preceding papers, our framework can be used to develop and evaluate different targeting policies. Second, we quantify the value of different types of information in the targeting of mobile ads, whereas the preceding papers were mainly concerned with click prediction.

Our paper also relates to the literature on ad networks' incentives to allow targeting. Levin and Milgrom (2010) were one of the first to conjecture the trade-off between value creation and market thickness. They argue that too much targeting can thin out markets, which, in turn, can soften competition and make the ad network worse off. This is particularly the case when there is significant heterogeneity in the distribution of advertisers' valuation of impressions (Celis et al. 2014). Building on this insight, a growing stream of analytical papers shows that there is a nonmonotonic pattern between the extent of targeting and ad network revenues (Bergemann and Bonatti 2011, Amaldoss et al. 2015, De Corniere and De Nijs 2016, Hummel and McAfee 2016, Sayedi 2018). A key difference between these papers and ours is that we do not make any distributional assumptions on the match values in our analytical model.

In spite of the increasing interest from the theoretical side, there has been limited empirical work on this topic with mixed findings. In an early paper, Yao and Mela (2011) present a structural model to estimate advertisers' valuations and show that targeting benefits both advertisers and the ad network. In a similar context, however, Athey and Nekipelov (2012) present a case study of two keywords and show that coarsening CTR predictions (worse targeting) can help a search advertising ad network generate more revenue. However, unlike our paper, neither of these papers can effectively design or evaluate counterfactual targeting regimes because their data come from highly targeted ecosystems without any randomization in ad allocation. More broadly, ours is the first empirical paper to view the revenue–efficiency trade-off through the lens of privacy and quantify the ad network's incentives to preserve users' privacy.

Next, our work relates to the literature on the interplay between privacy and targeting. Goldfarb and Tucker (2011b) use data from a series of regime changes in advertising regulations to show that restricting targeting reduces response rates and thereby advertisers' revenues. Similarly, Goldfarb and Tucker (2011a) and Tucker (2014) highlight the perils of excessive targeting because users perceive increased targeting as a threat to their privacy. Please see Goldfarb (2014) for an excellent review of targeting in online advertising and Acquisti et al. (2016) for a detailed discussion of consumer privacy issues. Our paper contributes to this literature by providing the first empirical evidence in support of the possibility of self-regulation in this market.

Finally, our paper adds to the growing literature on applications of machine learning in marketing, which focus on prediction problems; see Toubia et al. (2007) and Dzyabura and Yoganarasimhan (2018) for excellent summaries. Our paper contributes to this stream by demonstrating how a combination of theory-driven frameworks and machine learning methods can be used to go beyond prediction and help answer important substantive and prescriptive questions.

3. Setting and Data

3.1. Setting

Our data come from the leading mobile in-app advertising network of a large Asian country, which had over 85% market share in the category in 2015. The ad network works with over 10,000 apps and 250 advertisers, and it serves over 50 million ads per day (about 600 auctions per second). This ad network specializes in the Android operating system (OS). At the time of our study, smartphone penetration was reasonably high in the country, with over 60% of the population having access to smartphones. The share of the Android OS was over 85% of the market in this country in 2015, which is consistent with its share worldwide (Rosoff 2015).

3.1.1. Players. There are four key players in this marketplace.

Users. Individuals who use apps. They see the ads shown within the apps that they use and may choose to click on the ads.

Advertisers. Firms that show ads through the ad network. They design banner ads and specify their bid as the amount they are willing to pay per click and can include a maximum budget if they want to. Advertisers can target their ads based on the following variables: app category, province, connectivity type, time of day, mobile operators, and mobile brand of the impression. The ad network does not support more detailed targeting (e.g., behavioral targeting) at this point in time.

Publishers. App developers who have joined the ad network. They accrue revenues based on the clicks generated within their app. Publishers earn 70% of the cost of each click in their app (paid by the advertiser), and the remaining 30% is the ad network's commission.

Ad network or platform. It functions as the matchmaker between users, advertisers, and publishers. It runs a real-time auction for each impression generated by the participating apps and shows the winning ad during the impression. The platform uses a cost-per-click pricing mechanism and therefore generates revenues only when clicks occur.[2]

3.1.2. Auction Mechanism. The platform uses a quasi-proportional auction mechanism (Mirrokni et al. 2010). Unlike other commonly used auctions (e.g., second price or Vickrey), this auction uses a probabilistic allocation rule:

$$\pi_{ia} = \frac{b_a q_a}{\sum_{j \in \mathcal{A}_i} b_j q_j}, \tag{1}$$

where π_{ia} is the probability that advertiser a with bid b_a and quality score q_a wins impression i, and $\mathcal{A}_i$ denotes the set of advertisers participating in the auction for impression i. The quality score is an aggregate measure that reflects the advertiser's potential profitability for the platform. Currently, the platform does not use impression-specific quality scores; rather, it uses an advertiser-specific quality score that remained constant during our observation period.

Because of the probabilistic nature of the auction, the ad that generates the highest expected revenue for the platform is not guaranteed to win. Rather, advertiser a's probability of winning is proportional to $b_a q_a$.[3] Further, advertisers are only charged when a user clicks on their ad. The cost per click for an impression is determined using a next-price mechanism similar to that of Google's sponsored search auctions. In this case, the amount that the winning ad is charged per click is the minimum amount that guarantees its rank among the set of bidders. For example, suppose that there are three advertisers with bids 1, 2, and 3, and quality scores 0.1, 0.2, and 0.3, bidding on an impression. Then the products of bid and quality score for the three advertisers are 0.1, 0.4, and 0.9, respectively. In this case, if the second-ranked bidder wins the auction, he or she only needs to pay $1 \times 0.1/0.2 = 0.5$ because it is the minimum bid amount that guarantees that he or she will be ranked higher than the third-ranked bidder. Formally, we can write the cost per click for ad a in impression i as

$$\begin{aligned} CPC_{ia} = \inf\Bigg\{ b' \,\Bigg|\, & \sum_{j \in \mathcal{A}_i, j \neq a} \mathbb{1}(b' q_a \le b_j q_j) \\ & = \sum_{j \in \mathcal{A}_i, j \neq a} \mathbb{1}(b_a q_a \le b_j q_j) \Bigg\}, \end{aligned} \tag{2}$$

where $\sum_{j \in \mathcal{A}_i, j \neq a} \mathbb{1}(b_a q_a \le b_j q_j)$ is essentially the number of ads whose product of bid and quality score is lower than ad a, and the infimum over this set finds the minimum bid (b') that guarantees ad a's rank. Finally, note that the platform uses a fixed reserve price r_0 for all impressions. It is the minimum bid that is accepted by the platform. Thus, if an advertiser is not willing to pay at least r_0 per click, he or she is automatically out of competition.

3.2. Data

We have data on all the impressions and corresponding clicks (if any) in the platform for a 30-day period from September 30, 2015, to October 30, 2015. For each impression, we have data on

- *Time and date.* This is the time stamp of the impression.
- *Android advertising identification (AAID).* This is a user-resettable, unique device ID that is provided by the Android OS. It is accessible to advertisers and ad networks for tracking and targeting purposes. We use it as the user identifier in our main analysis.
- *App ID.* This is a unique identifier for apps that advertise through the platform.
- *Ad ID.* This is the identifier for ads shown in the platform.
- *Bid.* This is the bid that the advertiser has submitted for his or her ad; advertisers' bids do not change across impressions in our sample.
- *Cost per click (CPC).* This is the price that the winning advertiser has to pay if he or she wins the impression and a click occurs; this is calculated by the ad network based on Equation (2).
- *Location.* This includes the province as well as the exact location of a user, based on latitude and longitude.
- *Connectivity type.* This refers to the user's type of connectivity (e.g., Wi-Fi or cellular data).
- *Smartphone brand.* This is the brand of the user's smartphone (e.g., Samsung, Huawei).
- *MSP.* This is the user's mobile-phone service provider.
- *ISP.* This is the user's internet service provider.
- *Click indicator.* This variable indicates whether the user has clicked on the ad or not.

The total data we see in this one-month interval is quite large. Overall, we observe a total of 1,594,831,699 impressions and 14,373,293 clicks in this timeframe, implying a 0.90% CTR.

3.3. Data Splits and Sampling

We use the penultimate two days of our sample period (October 28 and 29) for training and validation and the last day for testing (October 30). We also use the preceding history from September 30 to October 27 (referred to as *global data*) to generate the features associated with these impressions. The splits of data are shown in Figure 1. Note that we do not fit our model on the global data because we do not have sufficient history to generate features for these impressions. Further, constraining all three data sets—training, validation, and testing—to a three-day window has advantages because recent research has shown that data freshness plays an important role in CTR prediction; that is, using older history for prediction can lead to poor predictive performance (He et al. 2014).

We draw a sample of 728,340 unique users (out of approximately 5 million) seen on October 28, 29, and 30 to form our training, validation, and test data sets.[4] In Online Appendix E.4, we show that this sample size is sufficient and that larger samples do not significantly improve model performance.

Figure 1 presents a visual depiction of the sampling procedure. Rows represent users. The impressions by users in our sample are shown using black points. There are 17,856,610 impressions in the training and validation data and 9,625,835 impressions in the test data. We have an additional 146,825,916 impressions by these users in the time preceding October 28, which form global data. These impressions will be used solely for feature generation (and not for model fitting). Note that both our user-based sampling procedure and feature-generation approach (see Online Appendix B) require us to be able to identify and track users. For this purpose, we use the *AAID* variable as our user identifier.

3.4. Summary Statistics

We now present some summary statistics on our training, validation, and test data, which constitute a total of 27,482,444 impressions. Table 1 shows the summary statistics of the categorical variables in the data. For each variable, we present the number of unique values, the share of top three values that the categorical variable can take, and the number of nonmissing data. Although we always have information on the app, ad, and time stamp of the impression, the other variables are sometimes missing. The shares are shown after excluding the missing variables in the respective category.

We observe a total of 263 unique ads and 9,709 unique apps in the data. The top three subcategories in each have large shares, and there is a long tail of smaller apps and ads. Moreover, as shown in Figure 2, we find that the top 37 ads account for more than 80% of the impressions, and similarly, the top 50 apps account for 80% of impressions.

Next, we present some descriptive analysis that examines the role of contextual and behavioral information in predicting CTR. A context is characterized by the when and where of an impression. As such, we define a unique context as a combination of an app and a specific hour of the day. Figure 3 shows the histogram of CTR for different contexts. As we can see, there is a significant amount of variation in CTR across contexts, which suggests that contextual information can be informative for predicting clicks. Next, to understand the role of behavioral information, we focus on the length of history available for a user. Figure 4 shows the cumulative distribution function (CDF) of the length of history for all the impressions and clicks. It suggests that users with longer histories are less sensitive to ads. Most of the clicks come from users with shorter histories, whereas most impressions come from users with longer histories. Thus, user-history or behavioral information

Figure 1. (Color online) Schema for Data Generation

Table 1. Summary Statistics for the Categorical Variables

		Share of top categories			
Variable	Number of categories	1st	2nd	3rd	Number of impressions
App	9,709	37.12%	13.56%	3.05%	27,482,444
Ad	263	18.89%	6.71%	6.31%	27,482,444
Hour of the day	24	7.39%	7.32%	6.90%	27,482,444
Province	31	25.25%	6.65%	6.51%	21,567,898
Smartphone brand	8	46.94%	32.30%	9.53%	25,270,463
Connectivity Type	2	54.64%	45.36%		27,482,444
ISP	9	68.03%	14.02%	7.09%	10,701,303
MSP	3	48.57%	43.67%	7.76%	26,051,042

also seems to be helpful in explaining the clicking behavior observed in the data.

4. Machine Learning Framework for Targeting

In this module, our goal is to develop a framework that can accurately estimate the gains in efficiency or the CTR for any targeting policy. To do this, we first need to specify and train a machine learning model that accurately predicts the match between an impression and an ad, that is, predicts whether an impression will generate a click or not, for both factual and counterfactual ads.

This section is organized as follows. We first define our problem in Section 4.1. Next, in Section 4.2, we discuss our empirical strategy. Here we explain the need for, and the extent of, randomization in our data-generating process and propose a filtering approach that establishes the scope of our framework in estimating both factual and counterfactual targeting policies. In Section 4.3, we present the details of our feature-generation framework. Finally, in Section 4.4, we discuss our estimation procedure, which consists of the learning algorithm, the loss function, and the validation method.

4.1. Problem Definition

Consider a setting with N impressions and A ads. We begin with a formal definition of a targeting policy.

Definition 1. A *targeting policy* τ is defined as a mapping between impressions to ads such that each impression is allocated one ad. For example, $\tau(i) = a$ means that targeting policy τ selects ad a to be shown in impression i.

In order to evaluate the effectiveness of a targeting policy, we first need an accurate prediction of CTR for each ad for a given impression in our data. That is, for each impression i and ad a, we need to estimate $\Pr(y_{i,a} = 1)$, where $y_{i,a}$ is the indicator that ad a receives a click when it is shown in impression i. This brings us to the formal definition of the match value matrix.

Definition 2. Let $m_{i,a} = \Pr(y_{i,a} = 1)$. The $N \times A$ match value matrix M is defined as

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,A} \\ m_{2,1} & m_{2,2} & \dots & m_{2,A} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N,1} & m_{N,2} & \dots & m_{N,A} \end{bmatrix}, \quad (3)$$

Figure 2. (Color online) Cumulative Fraction of Impressions Associated with the Top 100 Ads and Top 100 Apps

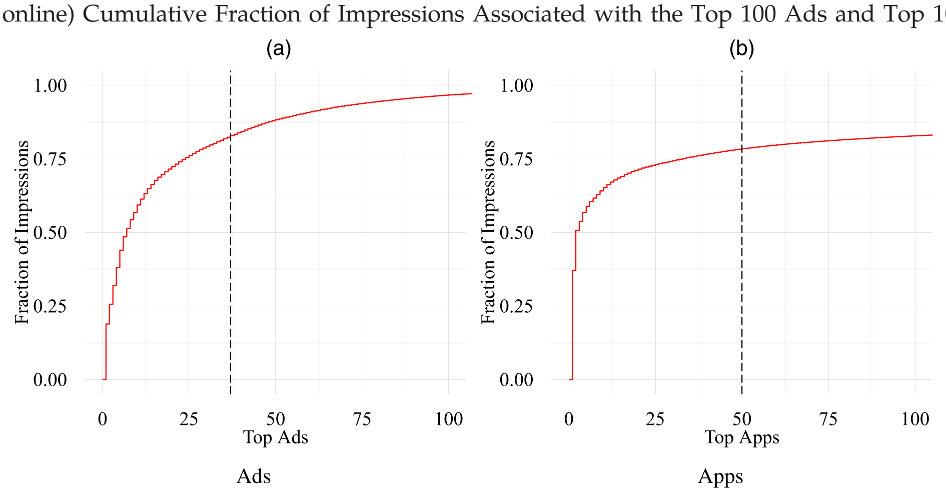
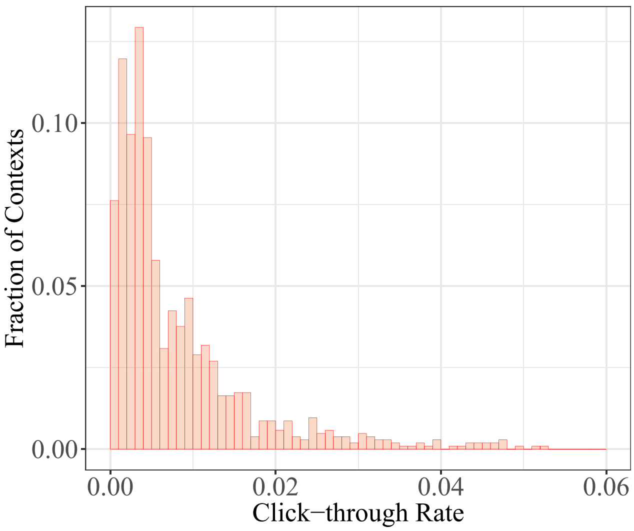


Figure 3. (Color online) Histogram of CTR for Different Contexts



Note. Context is defined as a unique combination of an app and an hour of the day (the where and when of an impression).

where N denotes the total number of impressions in our data, and A denotes the total number of ads competing for these impressions. There is a corresponding $N \times A$ matrix of outcomes Y, which consists of elements $y_{i,a}$. Note that we only observe the realized outcome for one element in each row or impression i for Y, which corresponds to the ad that was actually shown in that impression. The rest of the elements are treated as potential or unrealized outcomes.

In this section, our goal is to develop a machine-learning framework to estimate this match value matrix. We can use our estimated match value matrix $\hat{M}$ to perform the following analyses:

1. *Evaluate model performance.* We can evaluate the predictive performance of our model using the observed outcome. Let τ_0 denote the *current targeting policy* such that

$$\tau_0(i) = a_i, \tag{4}$$

where a_i is the ad that is actually shown in impression i. Because we observe the actual outcomes for y_{i,a_i}, we can evaluate how well our $\hat{m}_{i,a_i}$ estimates these outcomes.

2. *Evaluate the gains from efficient targeting policy.* Using the match value matrix, we can evaluate the expected CTR of any counterfactual targeting policy τ as

$$\hat{m}^{\tau} = \frac{1}{N}\sum_{i=1}^{N} \hat{m}_{i,\tau(i)}. \tag{5}$$

In particular, we are interested in the *efficient targeting policy* τ^* determined by our model that allocates each impression to the ad with the highest CTR for that impression:

$$\tau^*(i) = \underset{a}{\operatorname{argmax}}\, \hat{m}_{i,a}. \tag{6}$$

In Section 5.2, we quantify the gains in average CTR from efficient targeting over the current system.

4.2. Empirical Strategy

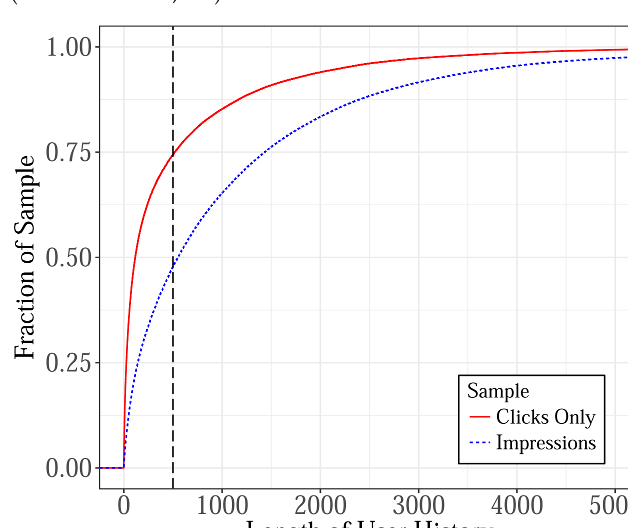
We now present our empirical strategy to estimate matrix M. At a high level, our goal is to build a model to predict whether an impression i showing ad a will receive a click or not based on the joint distribution of impressions and clicks in our data. That is, we seek to estimate a function $f(X_{i,a})$ such that

$$m_{i,a} = \Pr\big(y_{i,a} = 1\big) = f(X_{i,a}), \tag{7}$$

where $X_{i,a}$ is a set of features that are informative of whether impression i showing ad a will receive a click. Because this problem can be interpreted as function evaluation, we turn to machine learning algorithms that can capture complex relationships between the covariates and the outcome without imposing strong parametric restrictions on $f(\cdot)$.

Although machine learning methods can flexibly learn the function f from the data, their prediction power is bounded by the joint distribution of covariates and outcome (click) in the data. That is, these methods can accurately predict the outcome for an observation only if that observation could have been observed in the data. This requirement gives rise to

Figure 4. (Color online) Empirical CDF of the Length of User History for Impressions and Clicks (Truncated at 5,000)



Note. History is defined as the number of previous impressions from September 30 until the current impression.

two main challenges in evaluating counterfactual targeting policies.

Challenge 1. Function f cannot learn $m_{i,a}$ from the data if ad a could never have been shown in impression i; that is, ad a has zero propensity of being shown in impression i.

The reason is simple: if ad a could never have been shown in impression i, then the set of features $X_{i,a}$ is not within the joint distribution of the observed data. For example, if the ad for a fashion clothing brand was never shown in a sports app, then it is not possible to recover the fashion ad's click probability in the sports app.

It is worth noting that if the platform runs a deterministic auction (e.g., second-price auction), the set of ads that could have won the auction (and hence been shown during an impression) is a singleton. Similarly, the set of ads that can be shown in an impression in highly targeted environments would be very small. Therefore, data sets generated without any randomization in the ad-allocation mechanism will not allow researchers to push the scope of their analysis beyond the set of actual outcomes observed in the data. Randomization in ad allocation is thus necessary if we want to use our framework to evaluate the effectiveness of counterfactual targeting policies. This brings us to our first remark, which addresses Challenge 1.

Remark 1. Any ad participating in the auction for impression i ($\forall\, a \in \mathcal{A}_i$) has a nonzero propensity of being shown in impression i.

This is a direct result of the quasi-proportional auction run by the platform. As shown in Equation (1), each ad that participates in an auction has a nonzero probability of winning. This claim is the equivalent of the *positivity* or *overlap* assumption in the causal inference literature (Rosenbaum and Rubin 1983).

Although any kind of randomization can help overcome Challenge 1, we need to know the distribution of randomization to be able to correctly infer the click probability of counterfactual ads in any given impression i, that is, infer m_{ia} for ads $a \neq a_i$. If ads are randomized according to an unobserved rule, we may run into selection issues and obtain biased estimates of m_{ia}. We can characterize this challenge as follows.

Challenge 2. Function f cannot correctly infer match values (m_{ia}s) for counterfactual ads if the allocation rule is a function of an unobserved variable that is correlated with the outcome.

The following example helps illustrate this challenge: suppose that ad a_Y is targeted more toward younger users, whereas ad a_O is targeted more toward older users. Now, if younger users have a higher probability of click, failure to account for users' age will lead us to attribute the better performance of ad a_Y to the ad rather than to users' age. In the causal inference literature, this is usually known as *endogeneity* or *selection on unobservables* (Wooldridge 2010).

In our setting, we can simulate the allocation rule using the observed covariates. This gives us the unconfoundedness assumption, which we characterize in Remark 2.

Remark 2. For any impression i, ad allocation is independent of the set of the potential outcomes for participating ads ($a \in \mathcal{A}_i$), after controlling for the observed covariates. Thus,

$$\{y_{i,a}\}_{a\in\mathcal{A}_i} \perp\!\!\!\perp a_i \mid X_{i,a}. \tag{8}$$

Again, the allocation rule in Equation (1) directly satisfies the unconfoundedness assumption because everything on the right-hand side of this equation is known. First, for each i, we can infer the set of ads competing ($\mathcal{A}_i$) from our data because we observe all the targeting variables that can induce variation in $\mathcal{A}_i$. Second, advertisers do not change their bids, and the platform does not customize the quality score for each impression. Hence, $b_a q_a$ remains constant throughout our study, and we can easily infer propensity scores π_{ia} from the data, controlling for $\mathcal{A}_i$.

Together, in light of Remarks 1 and 2, we can estimate the match values $m_{i,a}$ not only for the ad that is shown in impression i but also for any counterfactual ad that could have been shown with nonzero propensity score. Naturally, estimates for small ads with very small probabilities of winning will be noisy. However, it is possible to overcome this issue by focusing on the top 37 ads that constitute over 80% of our data. In Section 4.2.1, we discuss our procedure for identifying the set of all participating ads in each impression that have nonzero propensity scores. Next, in Section 4.2.2, we discuss how we estimate these propensity scores and assess covariate balance.

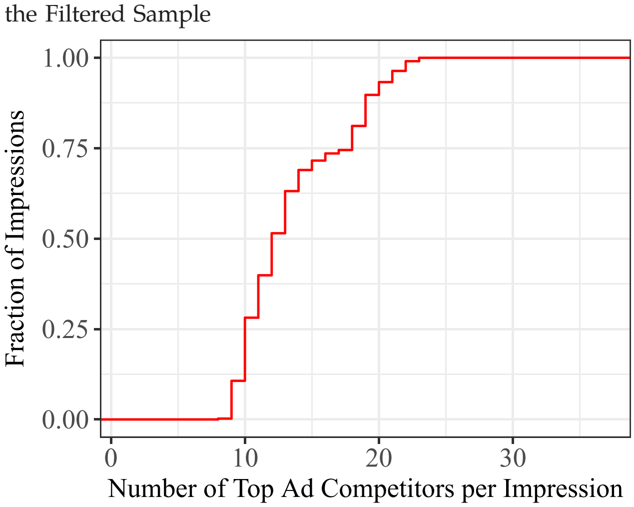
4.2.1. Filtering Procedure. As discussed earlier, if ad a could never have been shown in impression i, we cannot accurately estimate the match value for that impression–ad combination $m_{i,a}$. As such, we need to identify the set of participating ads in each impression and filter those that have zero propensity of being shown. In general, two factors influence whether an ad is available to participate in an auction for an impression.

- *Targeting.* Targeting by advertisers is the main reason why some ads are unavailable to compete for certain impressions and, therefore, have zero probability of being shown in them. For example, if an ad chooses to target only mornings, then it is not considered in the auctions for impressions in evenings. In this case, we should filter out this ad for all impressions in the evening. Although limited, targeting is nevertheless present in our setting and mainly happens on province, time, and app categories. Hence, for each impression i, we filter out all ads that were excluded from the auction for i because of targeting.

• *Campaign availability.* Second, some ads may be unavailable to compete for a given impression because their ad campaigns may not be running in the system when the impression happens. This could happen either because the advertiser's budget has been exhausted or because the advertiser has exited the market. Therefore, for each impression i, we filter out ads that were unavailable when it happens. Empirically, we find that campaign availability is not a major factor that leads to ad filtering because we focus on top ads.[5]

We now construct a filtering matrix $E_{N\times A} = [e_{i,a}]$ that filters out ads for each impression based on the factors just discussed, where each element $e_{i,a}$ takes value one if ad a has a nonzero probability of winning impression i and zero otherwise. Each row in this matrix shows which ads are competing for an impression.[6] However, our filtering may not be accurate for observations with missing targeting variables. Therefore, for all the analyses that use filtering, we focus only on the *filtered sample*, which consists of the impressions in the test data for which all targeting variables are nonmissing. Figure 5 shows the empirical CDF of the number of competing ads for each impression in the filtered sample data among the top 37 ads per impression. Note that almost all impressions have at least 8 top ads competing for it, and the median impression has 13 top ad competitors.

Figure 5. (Color online) Empirical CDF of the Number of Competitors (of the Top 37 Advertisers) per Impression for the Filtered Sample



4.2.2. Propensity Score Estimation and Covariate Balance.

As discussed earlier, the accuracy of our counterfactual match value estimates is predicated on the independence of assignment to ads and potential outcomes, given the observed covariates. Even though we know that this is theoretically true in our setting because of the allocation rule (Remark 2), we nevertheless need to empirically demonstrate the validity of this remark in our setting.

The standard practice in these cases is to assess and show covariate balance because it is a necessary condition for the unconfoundedness assumption. In simple settings, where both treatment and control are randomly assigned with a fixed probability to the entire population, we can easily assess balance by comparing the pretreatment variables across treatment and control groups. Our case is more complicated because of two reasons: (1) assignment to ads is not fully random but random given the propensity scores, and (2) because we focus on the top 37 ads, we have more than two treatment arms. To assess covariate balance, we therefore need to take the following steps:

1. *Propensity score estimation.* The first step is to estimate the propensity score π_{ia} for all a and i. Because we have multiple treatments, the dependent variable is a categorical variable with multiple classes. We use a multiclass XGBoost to estimate propensity scores given the success of machine learning methods in propensity score estimation (McCaffrey et al. 2013). Please see Online Appendix A.1 for more details.

2. *Assessing covariate balance.* Once we have the estimates for propensity scores, we can assess the balance for all pretreatment covariates. In our case, these covariates are the variables on which advertisers can target: province, app, time of day, smartphone brand, connectivity type, and MSP. To assess balance, we need to show that the inverse propensity-weighted distribution of each pretreatment variable is the same across all ads. Following the norm in the literature, we use the standardized difference between the population mean of a covariate and the inverse propensity-weighted mean of that covariate when assigned to ad a, and call it balanced if the difference is below 0.2 (McCaffrey et al. 2013). Please see Online Appendix A.2 for details on our balance measures and results.

4.3. Feature-Generation Framework

As discussed in Section 4.2, our goal is to build a model $m_{ia} = \Pr(y_{ia} = 1) = f(X_{ia})$ to accurately predict whether an impression i will receive a click or not. As such, we first need a vector of features $X_{i,a}$ that captures the factors that affect whether the user generating impression i will click on ad a.

It is important to generate an exhaustive and informative set of features because the predictive accuracy of our model will largely depend on the quality of the features we use. Given our research agenda, our features should also be able to capture the contextual and behavioral information associated with an impression over different lengths of history preceding the impression (long term, short term, and session level). To achieve these objectives, we adopt the main ideas from the functional feature-generation framework proposed by Yoganarasimhan (2020). There are three advantages to doing so. First, her function-based approach allows us to generate a large and varied set of features using a parsimonious set of functions. Second, it allows for a natural mapping between feature inputs and feature classification. Third, the general class of features she suggests has been shown to have good predictive power in this class of problems.

We now present a short overview of our feature functions and feature categorization, and we refer interested readers to Online Appendix B for a more detailed description.

4.3.1. Inputs for Feature Functions. To generate a set of features for each impression, we use feature functions that take some inputs at the impression level and output a corresponding feature for that impression. Our feature functions typically need two types of inputs:

- *Impression-specific information.* Each impression in our data can be uniquely characterized by three types of information, namely (1) contextual information that captures the context (where and when) of the impression (i.e., which app serves this impression and at what time (hour of day) is the impression being shown), (2) behavioral information that denotes the identity of the user generating this impression, and (3) ad-related information that denotes the identity of the ad that was shown during this impression.
- *History.* This input characterizes the history over which we aggregate to calculate the output of our functions. We define three different levels that capture the long-term (approximately one month), short-term (three days), and ongoing session-level history. Besides, we characterize the history in such a way that we can update the features in real time.

To reduce the dimensionality of our feature sets and boost the speed of our feature-generation framework, we group the smaller apps (below top 50) into one app category and all the smaller ads (below top 37) into one ad category. Thus, our features do not distinguish the context of smaller apps (ads) as separate from each other, though they are able to distinguish them from the top apps (ads). Please see Online Appendix B.1 for a complete formal definition of the inputs for feature functions.

4.3.2. Feature Functions. One challenge we face is that most of the information characterizing an impression–ad combination is categorical in nature, for example, the app showing the ad and the user seeing the ad. As a result, approaches that include all these categorical raw inputs and their interactions as covariates are prone to the curse of dimensionality. So we define functions that take these raw inputs as well as their interactions and map them onto a parsimonious set of features that reflect the outcome of interest—CTR.

We present an overview of our feature functions in Table 2 along with their functionality (see Online Appendix B.2 for a detailed description of the feature functions). These functions take different inputs based on the focal impression and return outputs that are integers or real numbers. These inputs are basically interactions of different raw inputs. The following examples give a high-level overview of what these functions do. Let p_i, t_i, u_i, and a_i denote the app,

Table 2. Feature Functions

Function	Functionality
Impressions	Number of impressions for a given set of inputs over a prespecified history
Clicks	Number of clicks for a given set of inputs over a prespecified history
CTR	Click-through rate for a given set of inputs over a prespecified history
AdCount	Number of distinct ads shown for a given set of inputs over a prespecified history
Entropy	Dispersion of ads shown for a given set of inputs over a prespecified history
AppCount	Number of distinct apps used by a given set of inputs over a prespecified history
TimeVariability	Variance in the user's CTR at different hours of the day over a prespecified history
AppVariability	Variance in the user's CTR across different apps over a prespecified history

hour, user, and ad associated with impression i. If the function *Impressions* is given p_i, u_i, and a_i and long-term history as inputs, it simply returns the number of times user u_i has seen ad a_i inside app p_i from the start of the data until the time at which impression i occurred. However, if it is only given u_i and short-term history, it returns the number of impressions user u_i has seen across all apps and ads over the last three days. Using this logic, we give different sets of inputs to these functions and generate 98 features for each impression i. In addition, we include a few standalone features such as dummies for each of the top ads, the user's mobile phone and internet service providers, latitude, longitude, and connectivity type. Overall, we have a total of 160 features for each impression–ad (*ia*) combination. Together, these features capture the interactive effects of advertising that are documented in the literature, such as carryover effects (Sahni 2015), spillover effects (Li and Kannan 2014), and effects of ad variety (Rafieian and Yoganarasimhan 2020). Please see Online Appendix B.3 for the full list of features.

4.3.3. Feature Categorization. All our features capture one or more type of information—contextual, behavioral, and ad specific. To aid our analysis, we therefore classify features based on the type of information used to generate them and group them into the following (partially overlapping) categories:

- *Contextual features* (F_C). These are features that contain information on the context of the impression app and/or hour of the day.
- *Behavioral features* (F_B). These are features that contain information on the behavior of the user who generated the impression.
- *Ad-specific features* (F_A). These are features that contain information on the ad shown during the impression.

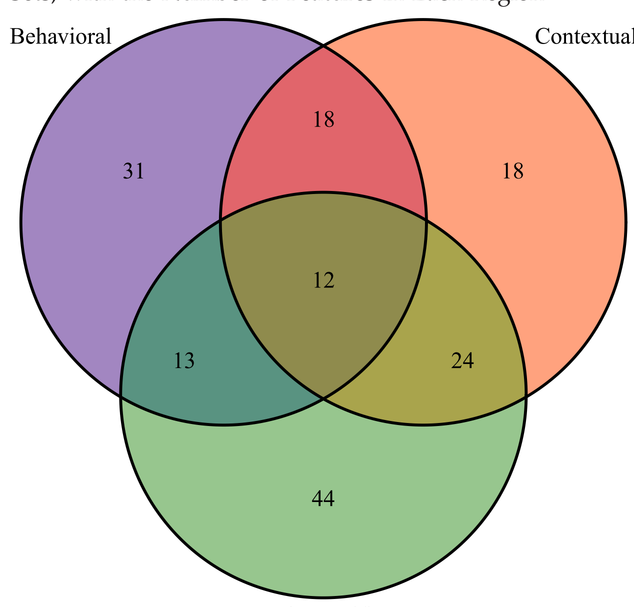
The three feature sets form our full set of features $F_F = F_B \cup F_C \cup F_A$. We now present a few examples of features generated using the *Clicks* function to elucidate this classification. The total clicks made by user u_i across all apps, ads, and hours of the day in the past month is a purely behavioral feature because it only contains information on the behavior of the user who generated impression i. In contrast, the total clicks made by user u_i in the app p_i over the last month constitute both a behavioral and contextual feature because it contains information on both the behavior of u_i and the context (app p_i) in which he or she made these clicks. Finally, the total clicks received by ad a_i over the last one month across all users, apps, and times is a purely ad-specific feature because it only reveals information about the ad's propensity to receive clicks. Thus, a feature can contain any combination of behavioral, contextual, or ad-specific information depending on the inputs used to generate it. Please see Table A1 in Online Appendix B for a mapping between each feature and the categories under which it falls and Figure 6 for a Venn diagram of our classification system.

4.4 Learning Algorithm: XGBoost

We now discuss the final step of our machine learning framework: the learning algorithm, which helps us learn the function $f(X_{i,a})$. It provides a mapping between our feature set ($X_{i,a}$) and the match value or click probability as $f(X_{i,a}) = m_{i,a} = \Pr(y_{i,a} = 1)$. Given that we want to maximize the predictive accuracy of the model, we do not want to impose parametric assumptions $f(\cdot)$. The problem of function evaluation is fundamentally different and harder than the standard approach used in the marketing literature, wherein we simply evaluate parameters after assuming a functional form. In the latter, the researcher only needs to search over the set of parameters given the functional form, whereas in the former we have to search over the space of functions. Therefore, we turn to machine learning algorithms that are designed for this task.

Specifically, we employ the XGBoost algorithm proposed by Chen and Guestrin (2016). XGBoost is a variant of the standard boosted regression trees and is one of the most successful prediction algorithms developed in the last few years. It has been widely adopted in both academia and industry.[7] At a high level, boosted regression trees can be thought of as performing gradient descent in function space using shallow trees as the underlying weak learners (Breiman 1998, Friedman 2001). Although boosted

Figure 6. (Color online) Venn Diagram of the Three Feature Sets, with the Number of Features in Each Region



trees have been around for over a decade, Chen and Guestrin's (2016) implementation is superior to earlier implementations from both methodological and implementation standpoints.[8] We refer interested readers to Online Appendix C for a more detailed description of XGBoost and now focus on two key components of our implementation: the loss function and the validation procedure.

To train any learning model, we need to specify how the model should penalize model fit, that is, the difference between the observed outcome y_{i,a_i} and model prediction $\hat{m}_{i,a_i}$ (where a_i refers to the ad shown in impression i). This is done using a loss function, which the machine learning algorithm minimizes. Because our outcome variable is binary, we use logarithmic loss (log loss) as our loss function. It is the most commonly used loss function in the CTR prediction literature (Yi et al. 2013) and has some attractive properties, for example, a faster convergence rate than other loss functions such as squared loss (Rosasco et al. 2004). The log loss for a model with predictions $\hat{\mathbf{M}}$ when the prediction matrix is Y can be written as

$$\mathcal{L}^{log\ loss}(\hat{M}, Y) = -\frac{1}{N}\sum_{i=1}^{N}\big(y_{i,a_i}\log(\hat{m}_{i,a_i}) + (1 - y_{i,a_i})\log(1 - \hat{m}_{i,a_i})\big). \tag{9}$$

Note that although the log-loss function takes as inputs the two matrices $\hat{M}$ and Y, the metric is calculated only over those ad–impression combinations that are actually observed in the data.

Validation is an important part of training any machine learning model. The boosting algorithm is designed to continuously update the prediction rule (or current estimate of $f(\cdot)$) to capture more and more complex relationships between the features $X_{i,a}$ in order to predict $y_{i,a}$. Because we do not impose any assumptions on the parametric form of $f(\cdot)$, this will likely lead to overfitting; that is, the model will evolve to fit too closely to the training data and perform poorly out of sample. Validation helps us avoid this problem by using parts of the data to validate the model. This ensures that the chosen model $f(\cdot)$ will have a good out-of-sample performance. Please see Online Appendix C.2 for a full description of our validation procedure.

5. Results from the Machine Learning Targeting Models

Recall that the goal of our machine learning framework is to estimate the matrix M defined in Equation (3). As such, our $\hat{M}$ contains CTR estimates for (1) the ads shown in the data and (2) counterfactual situations, that is, ads that could have been shown. In Section 5.1, we focus on the actual data and present results on the predictive performance of our framework on the observed sample. We also document the contribution of behavioral versus contextual information to our framework in this section. Next, in Section 5.2, we focus on the counterfactual estimates in $\hat{M}$ and evaluate the gains in CTR from an efficient targeting policy. Finally, in Section 5.3, we discuss robustness and scalability.

5.1. Predictive Performance of the Machine Learning Model

5.1.1. Evaluation Metric. To evaluate whether a targeting model improves our ability to predict clicks, we first need to define a measure of predictive accuracy or an evaluation metric. In line with our loss function, we use relative information gain (RIG), which is defined as the percentage improvement in log loss over the baseline that simply predicts average CTR for all impressions. Formally,

$$\mathrm{RIG}(\hat{M}, Y) = \left(1 - \frac{\mathcal{L}^{log\ loss}(\hat{M}, Y)}{\mathcal{L}^{log\ loss}(\bar{Y}, Y)}\right) \times 100, \tag{10}$$

where $\bar{Y}$ is an $N \times A$ matrix, each of whose elements is equal to $(\sum_{i=1}^{N} y_{i,a_i})/N$, that is, the average observed outcome of the sample or the average CTR of the data. Average CTR is the simplest aggregate metric available from any data, and using it as the baseline prediction tells us how well we can do without any model. It is important to control for this baseline because if the average CTR is very high (close to one) or very low (close to zero, as in most e-commerce settings, including ours), a naive prediction based on the average CTR leads to a pretty good log loss. Normalizing the log loss with the average CTR reduces the sensitivity of the metric to the data distribution (He et al. 2014). Nevertheless, we need to be careful when interpreting RIGs computed on different data sets because there is no obvious normalization in those cases (Yi et al. 2013).

In Online Appendix E.1, we present four other commonly used evaluation metrics: (1) mean squared error (MSE), (2) area under the curve (AUC), (3) 0/1 loss, and (4) confusion matrix. We discuss the pros/cons of these metrics and demonstrate the performance of our model on them.

5.1.2. Predictive Accuracy of the Full-Targeting Model. We now discuss our framework's ability to predict the actual outcomes in the data. Table 3 shows the gains in prediction for (1) training and validation data and (2) test data. The first row depicts the log loss for the Full model (which uses the set of all features and

Table 3. Log Loss and Relative Information Gain (RIG, in Percentage)

Evaluation metric	Training and validation	Test
Log loss for Full model	0.041927	0.044364
Log loss for baseline model	0.051425	0.054070
RIG of Full model	18.47%	17.95%

trains the XGBoost model). The second row depicts the log loss for the baseline model, which simply predicts the average CTR for the data set for all impressions. The third row is the RIG of the Full model compared with that of the baseline model.

The RIG of the Full model over the baseline is 17.95% on the test data, a substantial improvement in CTR prediction problems. This suggests that the data collected by the ad network is quite valuable and that our machine learning framework has significant predictive power on whether an impression–ad combination will receive a click.[9]

The RIG improvement for training and validation data is 18.47%, which is somewhat higher than the 17.95% for the test data. There are two potential reasons for this. First, all statistical models estimated on finite data have higher in-sample fit than out-of-sample fit. Indeed, this is the main reason we use the test data to evaluate model performance. Second, the difference could simply reflect the differences in the underlying data distributions for the two data sets. As discussed in Section 5.1.1, we cannot compare RIG across data sets because it is codetermined by the model and data. Thus, the difference between the RIG values across the data sets is not necessarily informative.

5.1.3. Value of Information: Behavioral vs. Contextual Features. We now examine the impact of different types of features on the predictive accuracy of our model. This is important for two reasons. First, data storage and processing costs vary across feature types. For example, some user-specific behavioral features require real-time updating, whereas pure contextual features tend to be more stable and can be updated less frequently. In order to decide whether to store and update a feature or not, we need to know its incremental value in improving targeting. Second, the privacy and policy implications of targeting depend on the features used. For example, models that use behavioral features are less privacy preserving than those that use purely contextual features. Before adopting models that are weaker on privacy, we need objective measures of whether such models actually perform better.

Recall that our features can be categorized into three broad overlapping sets: (1) behavioral, denoted by F_B, (2) contextual, denoted by F_C, and (3) ad specific, denoted by F_A. We now use this categorization to define two models:

- *Behavioral model.* This model is trained using behavioral and ad-specific features, without including any contextual features. Formally, the feature set used is $(F_B \cup F_A) \setminus F_C$.
- *Contextual model.* This model is trained using only contextual and ad-specific features, without including any behavioral features. The feature set for this model is $(F_C \cup F_A) \setminus F_B$.

Both models include ad-specific features that are neither behavioral nor contextual, for example, the total impressions received by the ad shown in the impression in the past month (Feature 2 in Table A1 in the Online Appendix B).[10] They also use the same loss function and training algorithm and only differ on the set of features used. Hence, it is possible for us to directly compare the RIG of one model over another within the same data.[11]

The results from these two models and their comparisons with the baseline model are presented in Table 4. First, consider the results for the full test data (presented in the second column). The Behavioral model has a 12.27% RIG over the baseline, which is considerably higher than 5.12%, the RIG of the Contextual model over the baseline. Together, these findings suggest that from a targeting efficiency perspective, behavioral information is more effective than contextual information in mobile in-app advertising.

This difference in the effectiveness of the two models directly relates to the extent of variation in the information used by the two models. The variation in behavioral features is much higher than the variation in contextual features because behavioral features are

Table 4. Comparison of Behavioral and Contextual Models for Different Samples of Test Data

Relative information gain over baseline	Full sample	Top ads and top apps	Filtered sample
Behavioral model	12.14%	14.82%	14.74%
Contextual model	5.25%	5.98%	6.77%
Full model	17.95%	22.85%	22.45%
No. of impressions	9,625,835	6,108,511	4,454,634
Percent of test data	100%	63.5%	46.28%

generated from the unique behaviors of over 700,000 users, whereas the total number of unique contexts is limited (to 1,200). Hence, the level of granularity of contextual features is much lower, and the Contextual model can only learn from aggregate outcome estimates across these limited contexts. Its ability to predict positive labels (i.e., clicks) is therefore much weaker than that of the Behavioral model.

One possible critique of the preceding analysis is that it does not exploit the full capacity of contextual information because we treat all the nontop ads as one advertiser category and all the nontop apps as one app category during feature generation (see Section 4.3.1). To address this issue, we consider a subsample of the test data that only consists of impressions that were shown in a top app and showed a top ad and rerun all the preceding comparisons. This accounts for 63.5% of our test data. The performance of our Full model on this subset of the data is even better than that on the full sample because there is no information loss on the ads or apps. The findings on the relative value of behavioral versus contextual features are even stronger in this data set, which suggests that our results in the full sample were not driven by the lack of good contextual information.

Finally, in the last column of Table 4, we show the performance of our model on the filtered sample (described in Section 4.2.1), which is the sample that we use for conducting our counterfactual analysis. Our qualitative findings remain the same for this sample too.

5.2. Counterfactual Analysis: Efficiency Gains from a CTR-Maximizing Targeting Policy

We now focus on an important counterfactual question from the platform's perspective: if the platform employs an efficient targeting policy such that each impression is allocated to the ad with the highest predicted CTR in that impression, to what extent can it improve the CTR in the system?

Recall that τ_0 and τ^* denote the current and efficient targeting policies, as defined in Equations (4) and (6), respectively. We can then use the following equation to calculate the gains in average CTR:

$$\rho(\tau^*, \tau_0; N_F) = \frac{\hat{m}^{\tau^*}}{\hat{m}^{\tau_0}} = \frac{\frac{1}{N_F}\sum_{i=1}^{N_F} \hat{m}_{i,\tau^*(i)}}{\frac{1}{N_F}\sum_{i=1}^{N_F} \hat{m}_{i,\tau_0(i)}}, \tag{11}$$

where N_F is the number of impressions in the filtered sample. It is crucial to conduct this counterfactual on the filtered sample (instead of the full sample) for the reasons discussed in Section 4.2.1.

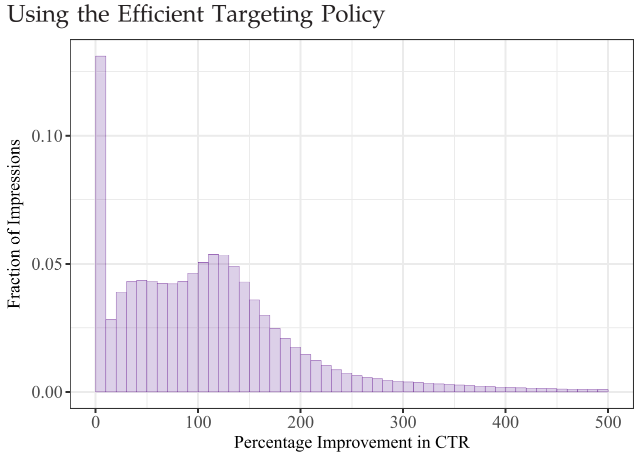
We find that an efficient targeting policy based on our machine learning model increases average CTR by 66.80% over the current regime. This is a substantial improvement and suggests that targeting based on behavioral and contextual features can lead to significant efficiency gains.

Next, we examine how efficiency gain varies by impression. Specifically, for each impression, we calculate the percentage improvement in CTR with efficient targeting as $(\hat{m}_{i,\tau^*(i)}/\hat{m}_{i,\tau_0(i)} - 1) \times 100$ and examine the distribution of this metric over impressions. In Figure 7, we show a histogram of this percentage improvement in CTR for the impressions in the filtered sample. We document considerable heterogeneity in CTR improvements across impressions: the median improvement in CTR is about 105.35%, implying that efficient targeting policy can make over half the impressions twice as clickable as the current system. The peak at the left side of the graph (at one) denotes cases where $\tau_0(i) = \tau^*(i)$, that is, where the platform happened to randomly select the ad that maximizes expected CTR.

This overlap between our efficient targeting policy and actual data allows us to evaluate the efficient targeting policy by inversely weighting the propensity scores for the actual outcomes in the overlapping area. This is a model-free approach known as *importance sampling*, which is commonly used in the policy evaluation literature (Dudík et al. 2014). We present the details of this approach in Online Appendix D and show that it establishes a 65.53% improvement in average CTR, which is similar to our findings based on Equation (11).

In sum, we find that an efficient targeting policy leads to significant gains in clicks for the platform using both model-based and model-free approaches. Nevertheless, a key question that remains unanswered is whether an efficient targeting policy is also revenue maximizing for the platform. Therefore, in Section 6, we incorporate competition and examine the relationship between efficiency and revenue.

Figure 7. (Color online) Histogram of Percentage Improvement in CTR over the Current System Using the Efficient Targeting Policy



5.3. Scalability and Robustness

We perform extensive checks on the robustness of all aspects of our machine learning approach and its scalability. We discuss these tests briefly here and refer readers to Online Appendix E for details.

First, in Online Appendix E.1, we show that our results are robust even if we use other evaluation metrics (AUC, MSE, 0/1 loss, and confusion matrix). Second, in Online Appendix E.2, we confirm that XGBoost is the best learning algorithm for our prediction task by comparing its performance with five other commonly used algorithms (least squares, least absolute shrinkage and selection operator (LASSO), logistic regression, classification and regression tree, and random forests). Third, in Online Appendix E.3, we run a few robustness checks on the feature-generation framework by considering alternative ways of aggregating over history as well as app-specific dummies. Again, we find no improvement in the model's predictive performance under these different specifications. Fourth, in Online Appendix E.4, we present some checks to establish that our data sample is sufficient and large enough to produce reliable results. Specifically, we find that the RIG gains start stabilizing with a sample of 100,000 users and that our sample of 728,340 users is more than sufficient for our purposes. Finally, in Online Appendix E.5, we show that our results are not sensitive to the validation procedure used to pick the tuning parameters by comparing with other methods, for example, hold-out validation and *k*-fold cross-validation.

6. Analysis of Revenue–Efficiency Trade-off

In Section 5, we showed that the ad network can substantially increase CTR with efficient targeting. However, that analysis was silent on the ad network's incentives to target and agnostic to revenues. In this section, we seek to answer two sets of important questions by focusing on competition and incentives. First, to what extent is the ad network incentivized to allow targeting, and is there an optimal level of targeting from its perspective? Second, how does the total surplus accrued by advertisers vary with targeting levels, and is there heterogeneity in advertisers' preferences on the optimal level of targeting?

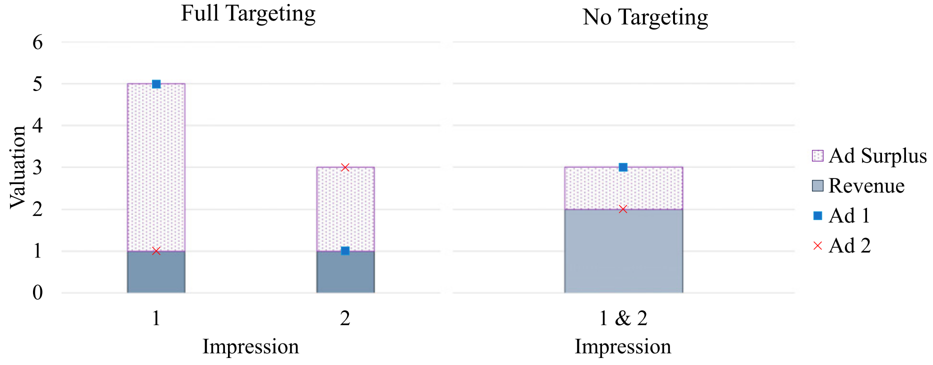
Incentives are particularly important in this context because if the platform is incentivized to not allow behaviorally targeted bids, then we may naturally converge to a regime with higher consumer privacy protection. In contrast, if the platform is incentivized to allow behavioral targeting, then an external agency (e.g., government) may have to impose privacy regulations that balance consumers' need for privacy with the platform's profitability motives. Similarly, if a substantial portion of advertisers prefer a more restrictive targeting regime, then the mobile ad industry can self-regulate. So we seek to quantify the platform's and advertisers' profits under different levels of targeting.

We now present an analytical framework to quantify the ad network's revenue–efficiency trade-off. This section proceeds as follows. In Section 6.1, we present a simple example to fix ideas and highlight the platform's efficiency–revenue trade-off. In Section 6.2, we present a stylized analytical model that characterizes the total surplus and platform revenues under different targeting strategies. In Section 6.3, we take this analytical model to data and present an empirical analysis of auctions with targeting.

6.1. A Simple Example

In an important paper, Levin and Milgrom (2010) argue that micro-level targeting can thin auction markets, which, in turn, can soften competition and make the platform worse off. In Figure 8, we present a

Figure 8. (Color online) Market Outcomes Under Full vs. No Targeting



Notes. The platform sells two impressions. Ad 1 and Ad 2 have valuations 5 and 1 for impression 1 and valuation 1 and 3 for impression 2, respectively. When bundled together, advertisers cannot distinguish between ads, giving an aggregate value of 3 and 2 to Ads 1 and 2, respectively. The entire shaded area in each case shows the total surplus generated. The area on the top is the share of advertisers and that on the bottom goes to the platform. See Online Appendix F for a detailed analysis of this example.

simple example to illustrate this idea. In this example, we consider a platform with two impressions and two advertisers whose valuations for these impressions do not align: advertiser 1 has a much higher valuation for impression 1 compared to impression 2, whereas the opposite is true for advertiser 2. Assume that the platform uses second-price auctions with cost-per-impression (CPI) pricing, where the highest bidder wins the impression and pays the bid of the second-highest bidder. We consider two regimes. In the full-targeting regime, the platform allows advertisers to submit targeted bids for each impression. In the no-targeting case, advertisers cannot distinguish between the two impressions and therefore have to submit the same bid for both the impressions (i.e., no targeted bidding). As shown in Figure 8, the platform cannot extract sufficient revenue if advertisers can distinguish between impressions (full targeting). However, the platform is able to extract more revenue by not revealing the identity of these impressions because advertisers are forced to rely on their aggregate valuation for both impressions together in this case. This example thus illustrates the platform's trade-off between value creation and value appropriation and highlights the platform's incentives to limit advertisers' ability to target.

6.2. Analytical Model of Auction with Targeting

We now develop a simple analytical model that captures the trade-offs discussed earlier. To reflect the idea of narrow targeting and thin markets, as envisioned by Levin and Milgrom (2010), we make two modeling choices. First, the idea that revenue loss in thin markets is due to the use of efficient auctions that guarantee that the highest valuation bidder will win (Krishna 2009). Although efficiency is satisfied in many auction mechanisms, we focus on second-price auctions because they are the most commonly used auctions in online advertising. Moreover, second-price auctions have the truth-telling property that makes our analysis more tractable. Second, the idea of narrow targeting by advertisers requires the pricing mechanism to be per impression. In a CPC mechanism, advertisers do not care about the match value of impressions because they are charged per click. For these reasons, we consider a setting where the platform uses a second-price auction mechanism with CPI pricing. Nevertheless, it is worth noting that neither of these two assumptions is essential to our analysis. Later in this section, we discuss how our results can be extended to other efficient auction mechanisms and/or CPC pricing.

As before, we consider a platform that receives N impressions and serves A advertisers. Let v_{ia} denote ad a's private valuation from impression i, and let V denote the value matrix

$$V = \begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,A} \\ v_{2,1} & v_{2,2} & \dots & v_{2,A} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N,1} & v_{N,2} & \dots & v_{N,A} \end{bmatrix}. \tag{12}$$

If an advertiser a can distinguish between all the impressions, he or she will submit targeted bids for each impression i. In a second-price auction, this is equivalent to a's valuation for impression i, $v_{i,a}$.

However, the extent to which advertisers can target depends on the level of targeting allowed by the platform. If certain information is not disclosed, advertisers may not be able to distinguish two impressions i and j. In such cases, a risk-neutral bidder's valuation for both impressions is the same and is equal to the expected value from the bundle of i and j (Sayedi 2018). For example, if the platform does not allow targeting at the app level, then advertisers cannot distinguish between impressions in two different apps, and their optimal strategy would be to submit the same bids for the impressions in both apps. Formally, we have the following definition.

Definition 3. Let I_l denote the set of impressions in bundle l. A targeting regime $\mathcal{I} = \{I_1, I_2, \dots, I_L\}$ denotes the platform's decision to bundle N impressions into L bundles such that advertisers can only bid for bundles and not impressions within the bundle. As such, impressions are only distinguishable across bundles, but not within a single bundle. That is, for bundle I_j, the advertiser a has the valuation $(\sum_{k \in I_j} v_{ka})/|I_j|$.

This definition characterizes all targeting regimes from impression-level targeting to no targeting. Impression-level targeting occurs when each impression is a bundle ($L = N$); that is, an advertiser can distinguish between all impressions and place targeted bids for each impression. By contrast, no targeting denotes the case where the platform bundles all impressions into one group ($L = 1$), implying that an advertiser can only have one valuation aggregated over all impressions ($\frac{1}{N}\sum_{i=1}^{N} v_{ia}$ for any a). Any intermediate strategy where $1 < L < N$ can be interpreted as partial targeting. An example of partial targeting is app-level targeting, where each bundle is an app, and impressions are distinguishable across apps but not within apps.

We can characterize the relative granularity of two targeting regimes as follows.

Definition 4. Let $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ denote two targeting regimes such that $\mathcal{I}^{(1)} = \{I_1^{(1)}, \dots, I_{L_1}^{(1)}\}$ and $\mathcal{I}^{(2)} = \{I_1^{(2)}, \dots, I_{L_2}^{(2)}\}$. Targeting regime $\mathcal{I}^{(1)}$ is at least as granular as $\mathcal{I}^{(2)}$

if, for any $I_j^{(1)} \in \mathcal{I}^{(1)}$, there exists a $I_k^{(2)} \in \mathcal{I}^{(2)}$ such that $I_j^{(1)} \subseteq I_k^{(2)}$. In words, if two impressions i and j are distinguishable in $\mathcal{I}^{(2)}$, then they will be distinguishable in $\mathcal{I}^{(1)}$.

We can use this definition to compare the granularity of two targeting regimes. For example, app-user-level targeting is more granular than app-level targeting. Now, the main question that the platform faces is at what level of granularity it should disclose information and allow targeting. Because we focus on the second-price auction, the highest-bidding ad in any impression wins that impression and pays the second-highest bid. This auction also guarantees the truth-telling property; that is, for each bundle, advertisers submit their aggregate valuation for that bundle as derived in Definition 3. The following proposition determines the relationship between the granularity level of targeting and market outcomes such as surplus and revenue.

Proposition 1. *Consider two targeting regimes $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ such that $\mathcal{I}^{(1)}$ is at least as granular as $\mathcal{I}^{(2)}$. Let $S^{(j)}$ and $R^{(j)}$ denote the total surplus and platform's revenue under targeting regime $j \in \{1, 2\}$. Then, for any distribution of valuations, $S^{(1)} \geq S^{(2)}$, but there is no fixed relationship between $R^{(1)}$ and $R^{(2)}$.*

Proof. See Online Appendix G.1. □

As the granularity of targeting increases, the total surplus generated increases, but the platform's revenue can go in either direction (unless we impose strong distributional assumptions on match values). Thus, although the matches are more efficient with more granular targeting, the platform may not be able to appropriate these efficiency gains. It is worth emphasizing that our analysis of revenue and surplus holds for any efficient auction because of the revenue equivalence theorem (Myerson 1981, Riley and Samuelson 1981). Further in Online Appendix H, we show that the same qualitative findings hold for a CPC pricing mechanism. Finally, note that this proposition is not applicable to a quasi-proportional auction because this is not an efficient mechanism.

6.3. Empirical Analysis of Auctions with Targeting

We now take this analytical model to data and examine market outcomes under different targeting regimes. Because the examination of the revenue–efficiency trade-off requires an efficient auction, our analytical model focuses on a second-price auction with a pay-per-impression payment scheme. However, notice that the mechanism in our data is a quasi-proportional auction with a pay-per-click payment scheme. Thus, our empirical analysis involves counterfactual evaluation of settings different from the one in our data.

As illustrated in our analytical model, the primary estimand that we require for our empirical analysis of auctions with different levels of targeting is matrix V defined in Equation (12). We can characterize each element in matrix V as follows:

$$v_{i,a} = v_a^{(c)} m_{i,a}, \tag{13}$$

where $v_a^{(c)}$ is the private valuation ad a gets from a click, and $m_{i,a}$ is the match valuations or expected CTR of ad a if shown in impression i.

In Section 6.3.1, we discuss how we can identify $v_{i,a}$ from our observed data. We then present our approach to obtain advertisers' click valuations in Section 6.3.2. Next, in Section 6.3.3, we explain how we can use our estimated match value matrix $\hat{M}$ from Section 4 to derive advertisers' match values under different targeting regimes. Finally, in Section 6.3.4, we discuss our empirical strategy to estimate the expected surplus, platform revenues, and advertisers' surplus.

6.3.1. Identification Strategy and Counterfactual Validity. To perform counterfactuals, we need to identify the elements of matrix V. Although we cannot directly identify $v_{i,a}$ from the data, we can separately identify both elements on the right-hand side of Equation (13)—the click valuation of ad a ($v_a^{(c)}$) and the match valuation of ad a when shown in impression i ($m_{i,a}$). To the extent that these two elements are policy-invariant primitives, our counterfactual analysis is valid. We now describe the basis on which these two estimands are identified and the conditions under which these are policy-invariant primitives.

- Click valuations are identified given advertisers' strategic bidding behavior in the current auction environment. The main assumption required is that advertisers select the bid that maximizes their utility. We can then specify advertisers' utility functions under the current auction observed in the data and use a first-order condition (FOC) to invert their observed bids to obtain consistent estimates of their click valuations. This is the standard identification strategy employed in the auction literature (Guerre et al. 2000, Athey and Haile 2007). It is worth noting that click valuations are policy invariant, although advertisers' bidding strategy can change under different auction mechanisms.
- Match valuations are identified given the unconfoundedness assumption: controlling for observed covariates, ad allocation is random. Please see Section 4.2 for a detailed discussion. Intuitively, match value estimates are policy invariant as long as users' underlying

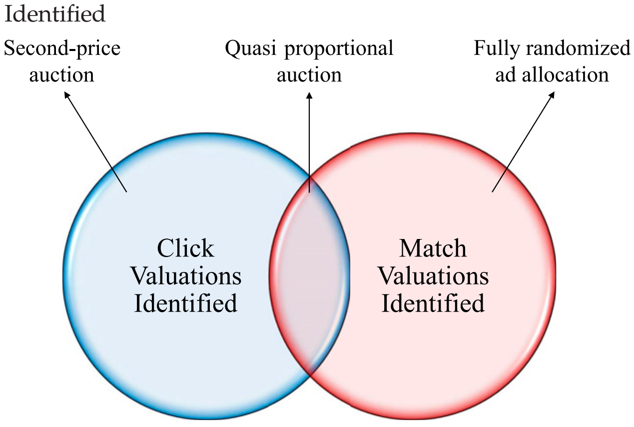
utility model for clicking on ads does not change under a different policy or auction.[12]

In sum, the identification of both click and match valuations is possible in settings that satisfy the unconfoundedness assumption while preserving the linkage between bids and click valuations. Figure 9 presents a Venn diagram of settings where each component is identified. It also highlights how common settings such as a second-price auction or a fully randomized ad allocation fail in this dual identification task. In standard auction mechanisms (e.g., second-price auctions), the identification problem stems from the deterministic allocation rule, which makes identification of match valuations impossible. In contrast, in a fully randomized experiment, there is no relationship between an advertiser's private click valuation and his or her observed bid, which makes the identification of click valuations impossible. To our knowledge, our setting (i.e., a quasi-proportional auction) is the *only* one in the literature that allows for the identification of both these components.

6.3.2. Estimation of Advertisers' Click Valuations. We now discuss the estimation of advertisers' click valuations $v_a^{(c)}$ based on the identification strategy discussed in Section 6.3.1. The standard approach in the structural auction literature is to assume that agents (advertisers in this case) are utility maximizing and derive the click valuations by inverting the equilibrium bidding function (Guerre et al. 2000, Athey and Haile 2007).

In our empirical setting, we observe that advertisers only submit one bid and do not change it (across impressions). Thus, we model advertiser a's bidding decision as a single-shot optimization, where he or she selects a bid b_a to maximize his or her own expected utility across all the impressions on which he or she bids. Let $\mathcal{G}_a$ denote advertiser a's beliefs about the joint distribution of the click valuations and quality scores of other advertisers bidding on the impressions for which a is competing.

Figure 9. (Color online) Venn Diagram Depicting Settings Where Click Valuations and Match Valuations Are Identified



Next, we define advertiser a's cost function as the expected payment that he or she has to make for each click that he or she receives, given bid b_a; that is, $c_a(b_a) = \mathbb{E}_{\mathcal{G}_a}[CPC_{ia}]$. Similarly, let $\pi_a(b_a)$ denote advertiser a's expected probability of winning an impression given bid b_a; that is, $\pi_a(b_a) = \mathbb{E}_{\mathcal{G}_a}[\pi_{ia}]$. Because the allocation function is proportional, we assume that $\pi_a(b_a) = b_a q_a/(b_a q_a + Q_{-a})$, where Q_{-a} is a constant reflecting the competitors' bids and quality scores.[13]

We can then characterize advertisers' equilibrium bidding strategy on our platform by taking the first order condition (FOC) of their expected utility. This FOC can then be inverted to obtain the click valuations, as shown in Proposition 2.

Proposition 2. *Consider a platform which runs quasi-proportional auctions where the allocation rule and CPC are given in Equations (1) and (2), respectively. Suppose that the cost function $c_a(b_a)$ is twice differentiable, $\{b_a^*\}_{a\in\mathcal{A}}$ is the set of observed bids, and $b_a^* c_a''(b_a^*)/c_a'(b_a^*) + 2 \geq 0$ for all ads. Then we can write the click valuation as*

$$v_a^{(c)} = c_a(b_a^*) + \frac{b_a^* c_a'(b_a^*)}{1 - \pi_a(b_a^*)}. \tag{14}$$

Proof. See Online Appendix G.2. □

We can obtain consistent estimates of click valuations from Equation (14) as long as we can observe/infer costs and bids from our data. We make three simplifications that make this task straightforward in our setting: (1) advertisers' probability of winning is close to zero (i.e., $\pi_a \approx 0$), (2) advertisers' CPC is approximately their bid (i.e., $c_a(b_a) \approx b_a$), and (3) the FOC in Equation (14) is satisfied for all advertisers, including reserve price bidders.[14] These three simplifications are reasonable in our empirical setting. First, as shown in Figure A.1 in Online Appendix A, even top ads that won the most impressions have a small probability of winning, justifying the first simplification. Second, on average, we find that an advertiser's CPC is over 92% of their bid, which provides support for the second simplification; that is, $c_a(b_a) \approx b_a$. Finally, the third simplification is also reasonable: only 11 of 37 ads are reserve price bidders.

With the three simplifications just outlined, Equation (14) can be approximated as

$$\hat{v}_a^{(c)} \approx 2b_a^*. \tag{15}$$

All the results presented in the main text are based on click valuations estimated based on Equation (15). However, in Online Appendix I.1, we present six alternative methods to estimate click valuations that progressively relax the simplifications made to derive Equation (15). Table A7 in Online Appendix I.1 presents an overview of the simplifications relaxed in each of the alternative methods. In particular, the last alternative method employs Rafieian's (2020b) recently proposed estimator for quasi-proportional auctions. His method is fully nonparametric and does not make any of the simplifications listed earlier. We find that the main results remain the same (qualitatively) even when we use these more complex estimators. Therefore, we stick with the simpler estimator in the main text and refer interested readers to Online Appendix I.1 for these robustness checks.[15]

6.3.3. Recovering Match Values. We now discuss how we can use our estimate of matrix M from our targeting framework to recover match values for any targeting regime. To start, $m_{i,a}$ is ad a's match value for any impression i if all impressions are distinguishable to him or her, and he or she is competing for that impression. This follows naturally from our arguments on the accuracy of match value estimates in Section 4.2. However, if two impressions are not distinguishable, the advertiser needs to use the aggregate estimate for that bundle. That is, for any targeting regime $\mathcal{I} = \{I_1, I_2, \ldots, I_L\}$, we can write the match value of advertiser a for impression i in bundle $\mathcal{I}$, $m^{\mathcal{I}}_{i,a}$, as follows:

$$\hat{m}^{\mathcal{I}}_{i,a} = \sum_{j=1}^{L} \mathbb{1}(i \in I_j) \frac{\sum_{k \in I_j} \hat{m}_{k,a} e_{k,a}}{\sum_{k \in I_j} e_{k,a}}, \quad \forall \; i \in \mathcal{I}, \tag{16}$$

where $e_{k,a}$ are elements of the filtering matrix that allows us to disregard inaccurate estimates and take the average of the rest. Figure 10 illustrates how the bundling and aggregation are performed on the match value matrix in a simple example with five impressions and three ads.

Here we assume that for any targeting regime $\mathcal{I}$, advertisers can infer their private match values for the bundles at that targeting regime $\hat{m}^{\mathcal{I}}_a$. This is reasonable because if the platform allows impression-level targeting, the platform would automatically share the impression-level data of each advertiser a with that advertiser (but not other advertisers). If a has sufficient data, then a can accurately estimate the match value vector $m_{i,a}$ for impression i from his or her own data. Similarly, if the platform only allows targeting at level $\mathcal{I}$, then advertisers would automatically have information on which bundle an impression belongs to as well as outcomes (whether impressions in a given bundle received clicks or not) and can therefore accurately infer their match values at the granularity of the bundle. Although this assumption always holds from a theoretical standpoint, it may not hold in practice because advertisers need sufficient data to obtain accurate estimates of their match values. Thus, the match value estimates of smaller advertisers and/or new advertisers can be noisy (though they will be consistent). In Online Appendix I.2, we show that our findings are robust even in situations where advertisers' match value estimates are noisy/imperfect.

Finally, the match value estimates derived from our quasi-proportional auction are assumed to remain the same under a second-price auction. This is reasonable because the match value simply indicates the click probability of a user in a given context for an ad. There is no economic rationale for users' click behavior to be a function of the auction, especially because users often do not know which auction is running on the back end. Intuitively, click and match valuations are treated as structural parameters.

6.3.4. Estimation of Revenue and Surplus. Given our estimates for click and match valuations, we can obtain estimates of the elements of the valuation matrix V as $\hat{v}_{i,a} = \hat{v}^{(c)}_a \hat{m}_{i,a}$. Further, we can estimate advertisers' expected value of impression i under targeting regime $\mathcal{I}$ as $\hat{v}^{\mathcal{I}}_{i,a} = \hat{v}^{(c)}_a \hat{m}^{\mathcal{I}}_{i,a}$. We now formally discuss our procedure to estimate revenue and surplus for any targeting regime $\mathcal{I}$.

First, we determine the winners of each impression as follows:

$$\hat{a}^*_i(\mathcal{I}) = \underset{a}{\operatorname{argmax}}\, \hat{v}^{\mathcal{I}}_{i,a} e_{i,a}, \tag{17}$$

Figure 10. (Color online) Construction of Match Value Matrix Under Targeting in a Simple Example with Five Impressions and Three Ads

Ad 1	Ad 2	Ad3
0.080	0.020	
0.060	0.060	0.090
0.080	0.010	0.150
0.160		0.100
0.120	0.050	0.050

Original match value matrix (empty cells are filtered)

Bundling

Ad 1	Ad 2	Ad3
0.080	0.020	
0.060	0.060	0.090
0.080	0.010	0.150
0.160		0.100
0.120	0.050	0.050

Impression bundling under targeting (each color corresponds to a bundle)

Aggregation

Ad 1	Ad 2	Ad3
0.070	0.040	
0.070	0.040	0.090
0.120	0.030	0.100
0.120		0.100
0.120	0.030	0.100

Match value matrix under targeting (values are aggregated within bundles) where $\hat{a}_i^*(\mathcal{I})$ is the winner for impression i under targeting regime $\mathcal{I}$. Note that the multiplication by the element of the filtering matrix $e_{i,a}$ simply ensures that the ad is competing in the auction for impression i and that the counterfactual match value estimates are valid, as discussed in Section 4.2.1.

Even though the winner is determined using the advertisers' expected value of impression i under a specific targeting regime, the surplus is calculated using the actual valuation matrix because it denotes the expected value that would be realized in the system if advertiser $\hat{a}_i^*(\mathcal{I})$ is allocated impression i. So we can write the surplus under targeting granularity $\mathcal{I}$ as

$$\hat{S}^{\mathcal{I}} = \sum_{i=1}^{N^F} \hat{v}_{i,\hat{a}_i^*(\mathcal{I})}. \tag{18}$$

To estimate the platform revenues, however, we need to use advertisers' expected values under targeting regime $\mathcal{I}$ because these values guide their bidding behavior. Further, we need to incorporate the fact that the revenue generated from impression i is the second-highest bid (or valuation) for it. Thus, the revenue under $\mathcal{I}$ is

$$\hat{R}^{\mathcal{I}} = \sum_{i=1}^{N^F} \max_{a \backslash \hat{a}_i^*(\mathcal{I})} \hat{v}_{i,a}^{\mathcal{I}} e_{i,a}. \tag{19}$$

Finally, we can estimate advertiser a's surplus under targeting regime $\mathcal{I}$ as follows:

$$\hat{W}_a^{\mathcal{I}} = \sum_{i=1}^{N^F} \left(\hat{v}_{i,\hat{a}_i^*(\mathcal{I})} - \max_{a \backslash \hat{a}_i^*(\mathcal{I})} \hat{v}_{i,a}^{\mathcal{I}} e_{i,a} \right) \mathbb{1}\left(\hat{a}_i^*(\mathcal{I}) = a\right). \tag{20}$$

This estimation is carried out on the filtered sample to ensure that our match value estimates are accurate, and hence, the averaging in the preceding equations is done over N_F. Figure 11 presents a step-by-step procedure to estimate revenue and surplus for the example case shown in Figure 10.

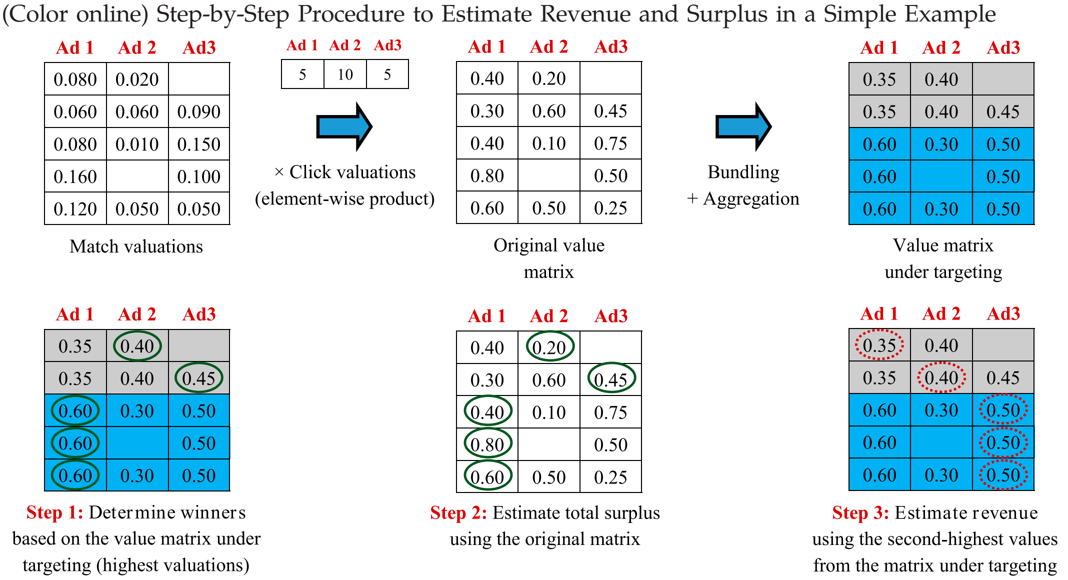
7. Counterfactual Results and Privacy Implications

Although we can analyze market outcomes for any targeting regime, we focus on the following four targeting regimes that have a one-to-one correspondence with our analysis in Section 5.1.3:

- *No targeting* $(\mathcal{I}^{(\mathcal{N})})$. The platform allows no targeting. As such, there is only one bundle (which constitutes all impressions), and advertisers cannot distinguish between any impressions.
- *Contextual targeting* $(\mathcal{I}^{(\mathcal{C})})$. The platform only allows contextual targeting. Here advertisers can distinguish between impressions in different contexts (app and time). However, impressions from different users in the same context are not distinguishable.
- *Behavioral targeting* $(\mathcal{I}^{(\mathcal{B})})$. The platform allows behavioral targeting, thereby allowing advertisers to distinguish between users but not contexts. Here advertisers can submit bids targeted at the user level but cannot distinguish two impressions by the same user in different contexts.
- *Full targeting* $(\mathcal{I}^{(\mathcal{F})})$. The platform allows impression-level targeting; that is, each impression is a bundle and therefore distinguishable. Advertisers can submit targeted bids for each impression.

Using Proposition 1, we can show that $S^{(\mathcal{F})} \geq S^{(\mathcal{N})}$ and that $S^{(\mathcal{C})}$ and $S^{(\mathcal{B})}$ lie in between because both contextual and behavioral targeting can be interpreted as imperfect targeting. However, we cannot theoretically pin down their relative magnitudes because these two types of information are orthogonal. (One cannot

Figure 11. (Color online) Step-by-Step Procedure to Estimate Revenue and Surplus in a Simple Example



be interpreted as being more granular than the other.) So we can only show that $S^{(\mathcal{F})} \geq S^{(\mathcal{C})}, S^{(\mathcal{B})} \geq S^{(\mathcal{N})}$. Further, we have no theoretical guidance on which of these targeting regimes maximizes platform revenues. We therefore use the empirical framework described in Section 6.3 to derive estimates of platform revenue, advertisers' surplus, and total surplus under the four targeting regimes, and we answer the question, "What is the optimal level of targeting that maximizes the platform's revenue?". The results from this exercise are shown in Table 5.

7.1. Platform's Revenue

Consistent with our theory model, our empirical results suggest that more granular targeting leads to higher efficiency in the market: the total surplus under full targeting is 13.02% higher than the no-targeting case. Further, in line with our findings in Section 5.1.3, we find that the total surplus under behavioral targeting is 2.13% higher than under contextual targeting. However, platform revenues exhibit more of an inverted-U-shaped curve. They are maximized when the platform restricts targeting to the contextual level. When the platform allows behavioral targeting, advertisers achieve a greater ability to target. Although this increases the total surplus in the system, much of this surplus is appropriated by advertisers, and the platform's revenue suffers.[16] Thus, the platform's incentives are not perfectly aligned with those of advertisers. Indeed, the platform's optimal targeting level is privacy preserving and aligned with consumers' preferences. We thus find support for the advertising industry's claim that the industry has natural economic incentives to limit user tracking/targeting and that self-regulation is feasible.

Our findings give rise to many interesting suggestions/ideas on optimal mechanism design and information revelation from the platform's perspective. Limiting targeting to the contextual level is an obvious strategy. However, this approach also reduces the total surplus and hence caps the platform's revenues. Thus, the optimal path for the platform may not be to restrict targeting but instead to consider mechanisms that can do both—increase efficiency and extract the revenue from winning advertisers by shrinking the informational rent. For instance, the platform could allow behavioral targeting and also adopt the standard theoretical solution proposed for revenue extraction—optimal reserve prices (Myerson 1981). Ostrovsky and Schwarz (2016) validate these theoretical findings using field experiments for search ads. However, they only consider optimal reserve prices for broad sets of keywords and assume CTRs to be homogeneous across advertisers. In contrast, we have a setting where each impression is a unique product, and advertisers' match values for an impression are heterogeneous. So, in our case, the platform has to develop a system that can set dynamic impression-specific optimal reserve prices.

7.2. Advertisers' Surplus

We begin by comparing total advertiser surplus across the four targeting regimes. As shown in Table 5, total advertisers' surplus is increasing with more granular targeting. This validates our theoretical prediction that more granular targeting helps advertisers by allowing them to generate more accurate estimates of their match values and place targeted bids. Under full targeting, advertisers' surplus is 11.69% of the total surplus, whereas this share drops to 0.69% when no targeting is allowed. Further, the share of advertisers' surplus under behavioral targeting is 8.99%, which is considerably higher than 6.13%, their share under contextual targeting. Together, these findings emphasize the value of behavioral information for advertisers.

Next, we explore whether all advertisers benefit when their ability to target is enhanced. In a competitive environment, greater ability to target does not necessarily translate into higher profits. Instead, it is the ability to target relative to competitors that matters. In Table 6, we show how many advertisers benefit as we move from one targeting regime (column) to another (row).

In general, more advertisers benefit when the platform allows more granular targeting, especially when it allows behavioral targeting. Moving from behavioral, contextual, and no targeting to full targeting benefits 23, 33, and 35 advertisers, respectively (first row of Table 6). However, more granular targeting is not uniformly better for all advertisers. The first

Table 5. Platform Revenues, Advertisers' Surplus, and Total Surplus for Different Levels of Targeting.

Targeting	Total surplus	Platform revenue	Advertisers' surplus
Full	9.45	8.35	1.10
Behavioral	9.18	8.35	0.84
Contextual	8.99	8.44	0.55
No targeting	8.36	8.30	0.06

Note. The numbers are reported in terms of the average monetary unit per impression.

Table 6. Number of Advertisers Who Benefit by Moving From One Targeting Regime (Column) to Another (Row) (Of 37 Top Advertisers).

To/from	Full	Behavioral	Contextual	Baseline
Full	NA	23	33	35
Behavioral	14	NA	34	36
Contextual	4	3	NA	33
Baseline	2	1	4	NA

Note. NA, not applicable.

column of Table 6 depicts situations where advertisers go from the most granular to less granular targeting regimes. Interestingly, it is populated with positive numbers, which suggest that some advertisers actually benefit from less granular targeting. For example, there are 14 advertisers who prefer behavioral targeting to full targeting. Similarly, although the majority of advertisers prefer behavioral targeting, there is a small portion of advertisers (3) who prefer contextual targeting. We present a simple example to highlight the intuition behind this: a nutrition supplement ad that advertises on a fitness app can get all the slots in that app at a low cost because other advertisers would place low bids when only app-level targeting is allowed. However, this ad would be worse off if only behavioral targeting is allowed because the competition for users in this app becomes more intense, and this ad will no longer be able to extract a large informational rent.

In sum, our findings offer some evidence that advertisers are likely to be differentially affected by privacy regulation on user tracking and behavioral targeting. Further research on the sources of heterogeneity in advertisers' incentives can help regulators craft the appropriate privacy policies.

7.3. Robustness Checks and Limitations

We run a series of robustness checks on the two main components of our estimation—click valuations and match valuations. First, in Online Appendix I.1, we consider alternative approaches to estimate click valuations from observed bids and show the robustness of our results. Second, in Online Appendix I.2, we show that our results are robust to the addition of noise to all the match value estimates (to reflect the cases where advertisers realize a noisy version of match value estimates from our machine learning framework).

Finally, although we have tried to make our analysis as exhaustive and complete as possible, our results should nevertheless be interpreted as short-run counterfactuals with the necessary caveats. First, we assume that advertisers' enhanced ability to target is only reflected in their targeted bidding. In reality, however, there might be value creation through other decision variables as well. Second, we assume that the set of ads competing for an impression will not change under different targeting regimes. This implies that there is no entry of new ads or exit of existing ads for an impression. Although this assumption may not be realistic, it is unlikely to change the qualitative findings of this paper. Third, we consider the case where the platform is a monopolist, which reflects our empirical setting. The question of how upstream competition affects privacy-preserving equilibrium outcomes is an important one, but outside the scope of our empirical setting. Finally, all our analysis is static. However, ad networks can adopt a forward-looking approach to allocate and sell ads. We refer readers to the recent series of work on adaptive ad sequencing that provides frameworks to maximize user engagement (Rafieian 2020a) and platform revenues (Rafieian 2020b).

8. Conclusions

Mobile in-app advertising is now a dominant ad format in the digital advertising ecosystem. In-app ads have unique tracking properties: they allow advertisers and ad networks to access the device ID of users' mobile devices and thereby enable high-quality behavioral targeting. Although this has made them appealing to advertisers, consumer privacy advocates are concerned about their invasiveness. Therefore, marketers and policymakers are interested in understanding the relative effectiveness of behavioral targeting compared to contextual targeting, the incentives of ad networks to engage in behavioral targeting, and the role of regulation in preserving privacy.

We propose a unified framework that consists of two components: a machine learning framework for targeting and an analytical framework for targeting counterfactuals when considering the competition in the market. We apply our framework to data from the leading in-app ad network of an Asian country. Our machine learning model achieves a RIG of 17.95% over the baseline when we evaluate it on test data. This translates to a 66.80% increase in the average CTR over the current system if we were to deploy an efficient targeting policy based on our machine learning framework. These gains mainly stem from behavioral information, and the value of contextual information is relatively small. Next, we build an analytical model of targeting and theoretically prove that although total surplus grows with more granular targeting between the ad network and advertisers, the ad network's revenues are nonmonotonic in the granularity of targeting. We then take our analytical model to data and conduct a series of targeting counterfactuals and show that the platform prefers to not allow behavioral targeting. There is also some heterogeneity among advertisers on their preferred level of targeting. Our findings suggest that ad networks have economic incentives to preserve users' privacy in the mobile advertising domain.

Our paper makes several contributions to the literature. First, from a methodological standpoint, we propose a unified framework for targeting that provides counterfactual estimates of platform revenues and efficiency under various targeting regimes. Our framework is generalizable and can be applied to a wide variety of advertising platforms as long as we are able to recover both match valuations and click valuations. In our setting, this is facilitated by the quasi-proportional auction, which induces randomness in the allocation of ads over impressions while preserving the linkage between observed bids and click valuations. However, other ad networks that employ deterministic auctions can also use our framework as long as they randomize ad allocation for a small portion of their traffic.[17] In such cases, (1) the data from the auctions can be used to recover click valuations, and (2) the data from the randomized traffic would satisfy the unconfoundedness assumption because of the exogenous variation in the allocation of ads and can be used to recover match valuations using our machine learning framework that combines ideas from causal inference and large-scale prediction tasks. Once these two primitives are available, our framework on revenue-efficiency analysis is directly applicable to evaluate market outcomes under different targeting scenarios.

Next, from a substantive perspective, our paper provides new insights on contextual and behavioral targeting. To our knowledge, this is the first paper to study both revenue and efficiency under these two types of targeting. Finally, from a policy point of view, we examine the incentives to target for the two major parties in the advertising ecosystem: the platform and advertisers. We expect our model and findings to speak to the debate on privacy regulations in the advertising industry.

Acknowledgments

The authors are grateful to an anonymous firm for providing the data and to the University of Washington–Foster High Performance Computing Laboratory for providing us with computing resources. The authors also thank Daria Dzyabura, Avi Goldfarb, Clarence Lee, Simha Mummalaneni, Puneet Manchanda, Sridhar Narayanan, Amin Sayedi, K. Sudhir, and Daniel Zantedeschi for detailed comments that have improved the paper. The authors also thank the participants of the 2016 Invitational Choice Symposium, the 2016 Federal Trade Commission and Marketing Science Conference, 2016 Big Data and Marketing Analytics Conference at the University of Chicago, the 2017 PhD Students Workshop at the University of Washington, the 2017 Adobe Data Science Symposium, the 2017 Summer Institute in Competitive Strategy, the 2018 Marketing Science Institute–Wharton Conference on New Perspectives in Analytics, the 2018 University of Washington-University of British Columbia conference, the 2019 University of Texas Dallas Frontiers of Research in Marketing Science conference, and Massachusetts Institute of Technology, Carnegie Mellon University, Harvard Business School, and Wharton marketing seminars for their feedback.

Endnotes

[1] Advertisers and ad networks have access to a unique device ID associated with the mobile device referred to as *IDFA* (ID for advertisers) in iOS devices and *AAID* (Android advertiser ID) in Android devices. This device ID is highly persistent and remains the same unless actively reset by the user.

[2] An impression lasts one minute. If a user continues using the app beyond one minute, it is treated as a new impression, and the platform runs a new auction to determine the next ad to show the user.

[3] From a practical perspective, probabilistic auctions ensure that individual users are not exposed to the same ad repeatedly within the same app session (which can be irritating). By contrast, in a deterministic auction, the same advertiser would win all the impressions until his or her budget runs out.

[4] Another approach would be to randomly sample impressions in each split of the data. However, this would not give us the complete user history for each impression in the training, validation, and test data sets. This, in turn, would lead to significant loss of accuracy in user-level features, especially because user history is sparse. By contrast, our user-based sampling approach gives us unbroken user history.

[5] Only six ads experience budget exhaustion (at least once) in the training data, four of which are entirely unavailable in the test data.

[6] Note that this information is not directly observed but inferred from advertisers' targeting decisions and campaign availability.

[7] Boosted trees in general, and XGBoost in particular, perform exceptionally well in tasks involving prediction of human behavior. Examples include store sales prediction, customer behavior prediction, product categorization, ad CTR prediction, and course dropout rate prediction. Indeed, almost all the Knowledge Discovery in Database (KDD) Cup winners have used XGBoost as their learning algorithm (either as a standalone model or in ensembles) since 2015.

[8] First, from a methodological standpoint, XGBoost can be interpreted as performing Newton boosting in the function space (as opposed to gradient descent) and thereby uses information from the Hessian as well. Thus, both the quality of the leaf structure and the leaf weights learned are more accurate in each step. Second, XGBoost uses a trick commonly used in random forests—column subsampling—which reduces the correlation between subsequent trees. Third, XGBoost employs a sparsity-aware split finding, which makes the algorithm run faster on sparse data. Finally, from an implementation perspective, XGBoost is highly parallelized, which makes it fast and scalable.

[9] One could argue that the significant predictive power of the Full model is due to the weak benchmark, which simply predicts average CTR for all impressions. Therefore, we also evaluate the performance of the Full model against two other baseline models: (1) ad-specific CTR and (2) targeting-area-specific CTR. The first model relates to ad networks' quality scoring practice; it predicts the average CTR for each ad as the match value for impressions showing that ad. The second model resembles the current targeting practice in the platform and predicts the average CTR for each targeting area (defined as the

intersection of all targeting variables) as the match value for all impressions within that targeting area. With these benchmark models as the denominator in Equation (10), we find that the Full model has a RIG of 16.86% over the ad-specific model and 10.06% over the targeting-area-specific model.

[10] We can also specify Behavioral and Contextual models that ignore ad-specific information. The qualitative results on the relative value of behavioral and contextual information for that case are similar to those presented here.

[11] As discussed in Section 5.1.1, RIG values are not directly comparable across different data sets. Simply put, in Table 4, comparisons within a column are interpretable, but comparisons across a row are not.

[12] Although we learn users' utility model flexibly using XGBoost without imposing a restrictive functional form on the utility function, we still require the underlying utility model to be policy invariant. This is equivalent to treating potential outcomes as structural parameters in the potential outcome framework (Imbens and Rubin 2015).

[13] Although there is no guarantee that $\pi_a(b_a)$ has the quasi-proportional form, it is easy to show by simulated experiments that it is a very accurate approximation. Further, advertisers know that the platform runs a quasi-proportional auction, so it is reasonable to assume that they rely on this functional form.

[14] The equality in Equation (14) may be invalid for reserve bidders because they may have submitted a reserve price bid because the platform did not allow them to submit a lower bid. Thus, in the presence of reserve price bidders, the distribution of bids that we see is truncated at the reserve price. In such a situation, we can only infer the truncated distribution of valuations. In Online Appendix I.1, we discuss how we can address this issue.

[15] The underlying theory behind our findings relates to match valuations and not click valuations: with more granular targeting, advertisers have more accurate match valuations, which, in turn, softens the competition and hurts platform's revenues. As such, the heterogeneity induced by allowing more granular targeting comes from the heterogeneity in match valuations, and click valuations are invariant to targeting scenarios. Therefore, using an approximate method to quantify the distribution of click valuations is sufficient for our purpose and does not change the main findings.

[16] Nevertheless, our findings are weaker than those predicted by theory models; that is, although revenues decrease with more granular targeting, the drop is not very large. This suggests that the strong distributional assumptions on the match values in earlier theory papers (e.g., Hummel and McAfee 2016) may not hold in real ad auctions.

[17] Indeed, this is standard practice in large ad networks; for example, Bing always randomizes ads for a small portion of its traffic (Ling et al. 2017). More broadly, all prominent ad networks now run A/B tests on portions of their data, and this portion of the traffic can be used to infer match valuations.

References

Acquisti A, Taylor C, Wagman L (2016) The economics of privacy. *J. Econom. Literature* 54(2):442–492.

Amaldoss W, Jerath K, Sayedi A (2015) Keyword management costs and broad match in sponsored search advertising. *Marketing Sci.* 35(2):259–274.

Athey S, Haile PA (2007) Nonparametric approaches to auctions. Heckman JJ, Leamer EE, eds. *Handbook of Econometrics*, vol. 6 (Elsevier, Amsterdam), 3847–3965.

Athey S, Nekipelov D (2012) A structural model of sponsored search advertising auctions. *Sixth Ad Auctions Workshop*, vol. 15.

Bergemann D, Bonatti A (2011) Targeting in advertising markets: Implications for offline vs. online media. *RAND J. Econom.* 42(3): 417–443.

Breiman L (1998) Arcing classifier. *Ann. Statist.* 26(3):801–849.

Celis LE, Lewis G, Mobius M, Nazerzadeh H (2014) Buy-it-now or take-a-chance: Price discrimination through randomized auctions. *Management Sci.* 60(12):2927–2948.

Chapelle O, Manavoglu E, Rosales R (2015) Simple and scalable response prediction for display advertising. *ACM Trans. Intelligent Systems Tech.* 5(4):61.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Krishnapuram B, ed. *Proc. 22nd ACM SIKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 785–794.

De Corniere A, De Nijs R (2016) Online advertising and privacy. *RAND J. Econom.* 47(1):48–72.

Dudík M, Erhan D, Langford J, Li L (2014) Doubly robust policy evaluation and optimization. *Statist. Sci.* 29(4):485–511.

Dzyabura D, Yoganarasimhan H (2018) Machine learning and marketing. Mizik N, Hanssens DM, eds. *Handbook of Marketing Analytics* (Edward Elgar Publishing, Northampton, MA).

Edwards J (2012) Apple has quietly started tracking iPhone users again, and it's tricky to opt out. *Bus. Insider* (October 11), http://www.businessinsider.com/ifa-apples-iphone-tracking-in-ios-6-2012-10.

Edwards-Levy A, Liebelson D (2017) Even Trump voters hate this bill he just signed. *Huffington Post* (April 3), https://www.huffpost.com/entry/trump-online-privacy-poll_n_58e295e7e4b0f4a923b0d94a.

Enberg J (2019) Digital ad spending 2019. *eMarketer* (March 28) https://content-na1.emarketer.com/us-digital-ad-spending-2019.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29(5):1189–1232.

Goldfarb A (2014) What is different about online advertising? *Rev. Indust. Organ.* 44(2):115–129.

Goldfarb A, Tucker C (2011a) Advertising bans and the substitutability of online and offline advertising. *J. Marketing Res.* 48(2): 207–227.

Goldfarb A, Tucker C (2011b) Online display advertising: Targeting and obtrusiveness. *Marketing Sci.* 30(3):389–404.

Guerre E, Perrigne I, Vuong Q (2000) Optimal nonparametric estimation of first-price auctions. *Econometrica* 68(3):525–574.

He X, Pan J, Jin O, Xu T, Liu B, Xu T, Shi Y, et al (2014) Practical lessons from predicting clicks on ads at Facebook. *Proc. 8th Internat. Workshop Data Mining Online Advertising,* (ACM, New York), 1–9.

Hummel P, McAfee RP (2016) When does improved targeting increase revenue? *ACM Trans. Econom. Comput.* 5(1):4.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, New York).

Kint J (2017) Opinion: Europe's strict new privacy rules are scary but right. *AdAge* (May 25), http://adage.com/article/digitalnext/europe-s-strict-privacy-rules-terrifying-apple/309155/.

Krishna V (2009) *Auction Theory* (Academic Press, Burlington, MA).

Levin J, Milgrom P (2010) Online advertising: Heterogeneity and conflation in market design. *Amer. Econom. Rev.* 100(2):603–607.

Li H, Kannan P (2014) Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *J. Marketing Res.* 51(1):40–56.

Ling X, Deng W, Gu C, Zhou H, Li C, Sun F (2017) Model ensemble for click prediction in Bing search ads. Barret R, ed. *Proc. 26th Internat. Conf. World Wide Web Companion* (ACM, New York), 689–698.

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF (2013) A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statist. Medicine* 32(19):3388–3414.

McMahan HB, Holt G, Sculley D, Young M, Ebner D, Grady J, Nie L, et al (2013) Ad click prediction: A view from the trenches. Dhillon I, ed. *Proc. 19th ACM SIGKDD Internat.*

Conf. Knowledge Discovery Data Mining (ACM, New York), 1222–1230.

Mirrokni V, Muthukrishnan S, Nadav U (2010) Quasi-proportional mechanisms: Prior-free revenue maximization. Lopez-Ortiz A, ed. *Latin Amer. Sympos. Theoret. Informatics* (Springer, Berlin), 565–576.

Myerson RB (1981) Optimal auction design. *Math. Oper. Res.* 6(1): 58–73.

Ostrovsky M, Schwarz M (2016) Reserve prices in internet advertising auctions: A field experiment. Working paper, Stanford Graduate School of Business, Stanford University, Stanford, CA.

Rafieian O (2020a) Optimizing user engagement through adaptive ad sequencing. Working paper, Cornell Tech and Cornell University, New York.

Rafieian O (2020b) Revenue-optimal dynamic auctions for adaptive ad sequencing. Working paper, Cornell Tech and Cornell University, New York.

Rafieian O, Yoganarasimhan H (2020) How does variety of previous ads influence consumers ad response? Working paper, Cornell Tech and Cornell University, New York.

Riley JG, Samuelson WF (1981) Optimal auctions. *Amer. Econom. Rev.* 71(3):381–392.

Rosasco L, De Vito E, Caponnetto A, Piana M, Verri A (2004) Are loss functions all the same? *Neural Comput.* 16(5):1063–1076.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rosoff M (2015) The research firm that once thought Microsoft would beat the iPhone has given up on Windows Phone. *Bus. Insider* (December 7), https://www.businessinsider.com/idc-smartphone-os-market-share-2015-12.

Sahni NS (2015) Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quant. Marketing Econom.* 13(3):203–247.

Sayedi A (2018) Real-time bidding in online display advertising. *Marketing Sci.* 37(4): 553–568.

Toubia O, Evgeniou T, Hauser J (2007) Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design. Gustafsso A, Herrmann A, Huber F, eds., *Conjoint Measurement: Methods and Applications* (Springer, New York), 231–258.

Tucker CE (2014) Social networks, personalized advertising, and privacy controls. *J. Marketing Res.* 51(5):546–562.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Yao S, Mela CF (2011) A dynamic model of sponsored search advertising. *Marketing Sci.* 30(3):447–468.

Yi J, Chen Y, Li J, Sett S, Yan TW (2013) Predictive model performance: Offline and online evaluations. Dhillon I, ed. *Proc. 19th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 1294–1302.

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Sci.* 66(3):1045–1070.