Should We Share Our Cookies? The Effect of Data on Match Quality in Online Advertising

Mehdi Akhbari[1,*], Farhang Shamsoddin[2]

Abstract

User data plays a central role in online advertising by enabling platforms to match ads to users more effectively. This paper quantifies how access to additional data affects match quality, measured by user click-through rates, and evaluates how data-driven mergers between platforms with complementary datasets influence market outcomes. Mergers improve match quality through two distinct mechanisms: (i) a precision effect, whereby a larger number of observations improves the accuracy of predicted click probabilities, and (ii) a targeting effect, whereby broader tracking of user behavior enhances the platform's ability to tailor ads to user preferences. The analysis utilizes the conditionally random allocation of ads in the data provided by Iran's largest online advertising platform and employs causal forest, a novel causal machine learning method, to estimate heterogeneous user click rates based on extensive browsing and ad exposure data from more than 1.6 million internet users. Simulations of user and platform behavior reveal that hypothetical mergers could boost match quality by 14-25%. In addition, this research finds that the majority of gains arise from an increase in the number of observations while expanding the set of observable user characteristics provides comparatively smaller gains.

*corresponding author

Email addresses: akhbari@utexas.edu (Mehdi Akhbari), farhang.shamsoddin@mail.utoronto.ca (Farhang Shamsoddin)

[1]Depratment of Economics, The University of Texas at Austin, 2225 Speedway, Austin, Texas, United States

[2]Rotman School of Management, University of Toronto, 105 St. George St., Toronto, Canada

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Should We Share Our Cookies? The Effect of Data on Match Quality in Online Advertising

Abstract

User data plays a central role in online advertising by enabling platforms to match ads to users more effectively. This paper quantifies how access to additional data affects match quality, measured by user click-through rates, and evaluates how data-driven mergers between platforms with complementary datasets influence market outcomes. Mergers improve match quality through two distinct mechanisms: (i) a precision effect, whereby a larger number of observations improves the accuracy of predicted click probabilities, and (ii) a targeting effect, whereby broader tracking of user behavior enhances the platform's ability to tailor ads to user preferences. The analysis utilizes the conditionally random allocation of ads in the data provided by Iran's largest online advertising platform and employs causal forest, a novel causal machine learning method, to estimate heterogeneous user click rates based on extensive browsing and ad exposure data from more than 1.6 million internet users. Simulations of user and platform behavior reveal that hypothetical mergers could boost match quality by 14-25%. In addition, this research finds that the majority of gains arise from an increase in the number of observations while expanding the set of observable user characteristics provides comparatively smaller gains.

Keywords: Online Advertising, E-commerce platforms, Data sharing, Causal inference, Causal machine learning, Platform economics, Digital markets, Big data, Data economy, Advertisement Targeting, Personalization, Antitrust

1. Introduction

Data plays a central role in digital platform markets by enabling firms to better predict user preferences and personalize products. In online platforms,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

firms exchange services for user data and attention, which can be used for prediction, pricing, and product customization.

Personalization is a key mechanism through which data improves match quality in markets with heterogeneous preferences, such as digital advertising and online platforms. By analyzing user behavior, platforms predict preferences and tailor content and advertisements accordingly.

A primary source of data for these platforms is user interaction. As users browse content and respond to ads, they reveal information about their preferences, allowing platforms to improve targeting and increase engagement. This feedback loop creates strong incentives to acquire additional data, often through mergers with firms holding complementary datasets. Prominent examples include Google–YouTube, Facebook–Instagram, and Microsoft–LinkedIn.

At the same time, increased data concentration has raised antitrust concerns. Large digital platforms may leverage their data advantage to reinforce market power, prompting calls for regulatory interventions, including potential breakups.

Despite growing policy interest, there is limited empirical evidence on how data affects market efficiency and match quality in digital markets. This paper quantifies the impact of data on match quality in the online advertising industry and evaluates how mergers between platforms with complementary data affect market outcomes.

A data-driven merger can improve match quality through two mechanisms. First, a *Precision Effect*: access to more observations improves the accuracy of predicted click probabilities, leading to better ad selection. Second, a *Targeting Effect*: broader tracking of user behavior allows the platform to better capture preference heterogeneity and tailor ads more effectively. We disentangle and quantify the contribution of each mechanism.

We use a novel dataset from the largest online advertising platform in Iran, which features a conditionally random ad allocation mechanism. This institutional feature provides exogenous variation in ad exposure, allowing us to estimate heterogeneous user click probabilities across ads and users. We then use these estimates to simulate counterfactual market outcomes under alternative data-sharing regimes. To estimate heterogeneous effects, we employ causal forest methods.

The results show that using all available data more than doubles average click rates relative to the status quo. Moreover, data-driven mergers increase match quality by 14–25%, while breaking up a monopolistic platform reduces match quality by 12–20%. Decomposition analysis reveals that most of the gains arise from improved statistical precision, with only 18–25% attributable

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

to increased observability of user characteristics.

This paper makes three main contributions. First, it provides causal evidence on how access to data affects match quality in a large-scale digital advertising market, leveraging quasi-random ad allocation to identify heterogeneous user responses. Second, it decomposes the gains from data into a precision effect, driven by increased sample size, and a targeting effect, driven by richer user observables, showing that the former accounts for the majority of improvements in performance. Third, it quantifies the efficiency implications of data-driven mergers and platform breakups by simulating counterfactual market structures, providing empirical guidance for antitrust policy in data-intensive markets.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the theoretical framework and describes the platform environment and counterfactual market structures. Section 4 describes the data. Section 5 outlines the empirical model. Section 6 reports the estimation results and documents heterogeneity in click behavior. Section 7 presents the counterfactual simulations and quantifies the effects of data on match quality. Section 8 concludes.

2. Related Literature

This paper contributes to several strands of economic literature. The first one is data economy literature, which models data as an input into firms' production processes (e.g., Farboodi and Veldkamp (2021), Jones and Tonetti (2020)). A central insight from this literature is the presence of a "data feedback loop," whereby firms generate data through production and use that data to improve productivity, leading to self-reinforcing growth dynamics.

This paper provides causal evidence on the relationship between data and productivity. This relationship is crucial because the extent to which a firm's productivity depends on data determines the long-term market structure. If the returns to data are increasing, firms that start out larger can generate more data, boost productivity, grow rapidly, and become superstars Farboodi and Veldkamp (2021). Conversely, if the returns to data decrease (as expected for larger data volumes), smaller firms with similar initial productivity levels can eventually converge in size. A key challenge in this setting is reverse causality: more productive firms generate more data, and more data increases productivity. We address this challenge using exogenous variation in ad exposure generated by a conditionally random allocation mechanism.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Empirical evidence on the value of data remains limited. Bajari et al. (2019) show that the number of observations improves demand forecasting on Amazon, while Tambe et al. (2020) find that digital capital predicts firm productivity. While none of these papers claim to find causal relationships, Wernerfelt et al. (2024) use a randomized experiment to show that "offsite" tracking data reduces advertiser costs. While this research identifies a causal relationship and demonstrates the value of data, it does not examine how variations in the amount of data affect match quality.

This paper is also related to the literature on data and market power. Data accumulation can generate economies of scale and increase market concentration in digital platforms, creating a trade-off between efficiency and market power (Cremer et al. (2023)). While a large theoretical literature, such as Ali et al. (2023); Brynjolfsson and Mitchell (2017); Prat and Valletti (2022); Calvano et al. (2023); Hagiu and Wright (2023); Chen et al. (2022), studies this trade-off, empirical work has primarily focused on price effects. For example, Shiller et al. (2013) and Dubé and Misra (2017) show that price discrimination increases platform profits, while Goldfarb and Tucker (2019) and Shiller and Waldfogel (2009) document deviations from optimal pricing.

In contrast to this literature, this paper focuses on the efficiency effects of data in a market setting. To the best of our knowledge, it is the first empirical study to quantify how mergers affect match quality in a data-intensive industry by expanding access to user-level data.

3. Theoretical Framework

This paper studies the online advertising market in Iran using data from the country's largest advertising platform. The platform operates across more than 2,300 websites, serves billions of ad impressions per month, and reaches a large share of internet users. Due to limited presence of international competitors, it plays a central role in the domestic digital advertising ecosystem.

3.1. Platform's Role at the Status Quo

The platform matches advertisers with users visiting partner websites. When a user visits a website, an ad request is sent to the platform, which selects ads from a pool of advertisers and displays them. Advertisers pay on a per-click basis, and revenue is shared between the platform and the website.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

3.1.1. Ad Allocation Mechanism

The platform uses a proportional allocation rule (Mirrokni et al. (2010)) in which ads are displayed with probabilities that depend on bids and historical performance:

$$e_{awt} = \frac{\alpha_{awt} b_{awt} CTR_{awt}}{\sum_{a' \in A_{wt}} \alpha_{awt} b_{a'wt} CTR_{a'wt}} \tag{1}$$

where e_{awt} is the probability of ad a being shown on website w at time t (also denoted as *propensity score* of ad a on website w at t), b_{awt} is the (per-click) bid submitted for ad a on website w at t, CTR_{awt} is the historical average click-through rate for ad a on website w by time t, and A_{wt} is the set of all advertisers advertising on website w in that time. Finally, α_{awt} is a multiplying coefficient that affects the propensity scores over time. It is used to boost propensity scores for new ads or lower them for ads approaching their daily budget limits.

Importantly, this mechanism assigns a positive probability to all ads on a given website, generating quasi-random variation in ad exposure. In particular, conditional on the website visited and the timing of the impression, the remaining variation in ad exposure is effectively random. This feature is crucial for identifying heterogeneous user responses to ads.

3.2. Counterfactuals

This paper studies two counterfactual market structures: (i) a monopoly in which a single platform has access to all data, and (ii) a duopoly in which data is split across platforms. In both cases, platforms allocate ads to maximize their expected revenue. Since expected revenues are functions of the expected click rates, which are predicted using the platform information sets, the ads shown to each user may be different under different market and information structures. In addition, the set of potential ads, the websites users visit, and the number of ad slots per page are assumed to remain fixed and equal to those observed in the status quo across all counterfactuals.

3.2.1. Monopoly

In the monopoly counterfactual, the platform uses all available user data to predict click probabilities and allocates ads to maximize expected revenue. Compared to the status quo, the key difference is that the platform fully exploits user browsing history and past ad exposure to improve targeting. This counterfactual is depicted in Figure 1.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

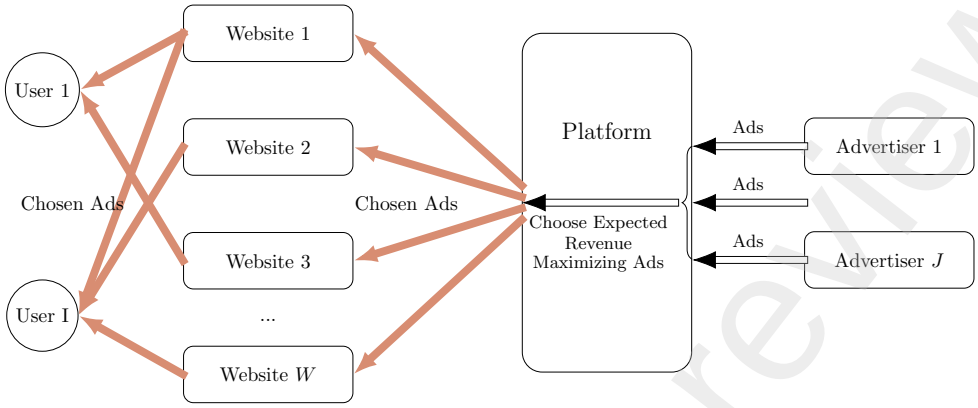


Figure 1: Simplified Scheme of the Platform's Role at the Monopoly Counterfactual

3.2.2. Duopoly

In the duopoly counterfactual, two platforms operate on disjoint subsets of data, each observing user behavior only on the websites they serve. This fragmentation reduces the information available for predicting user preferences and leads to less efficient ad targeting. We consider several intuitive data splits, including random partitioning by website, website category, and users. These duopoly counterfactuals are discussed in more detail in Section 7.3

Figure 2 illustrates an example of these counterfactuals. In this case, websites $1, 2, ..., W_1$ use Platform 1, and websites $W_2, ..., W$ partner with Platform 2. Thus, when Platform 1 determines which ads to show to user 1, it does not know that this user has visited website W_2, what ads she observed on that website, and what her response was. If any of those observables can provide additional information about her interest in different ads, Platform 1 may be unable to show her the most relevant ads, leading to a lower match quality relative to the monopoly counterfactual.

3.3. Sources of Match Quality Gains

In this setting, two distinct mechanisms can increase the match quality after a merger:

1. *Precision Effect*: Access to more observations improves the accuracy of estimated click probabilities and reduces the likelihood of selecting suboptimal ads due to sampling noise.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

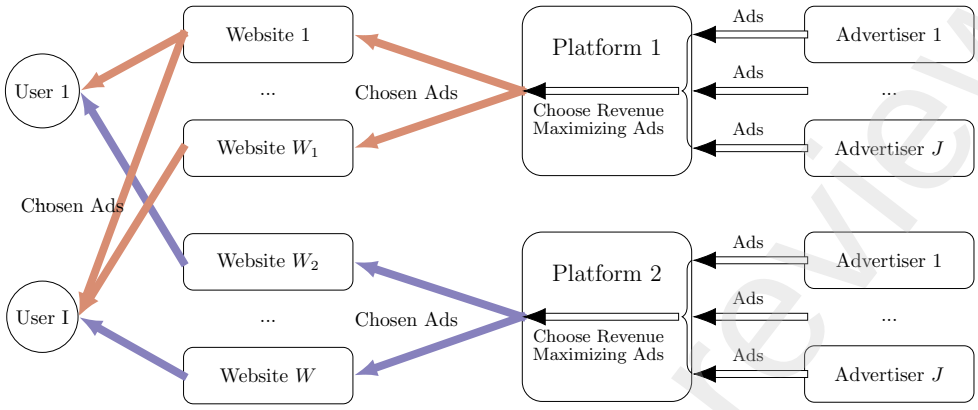


Figure 2: Simplified Scheme of the Platform's Role at a Duopoly Counterfactual (Split by Website)

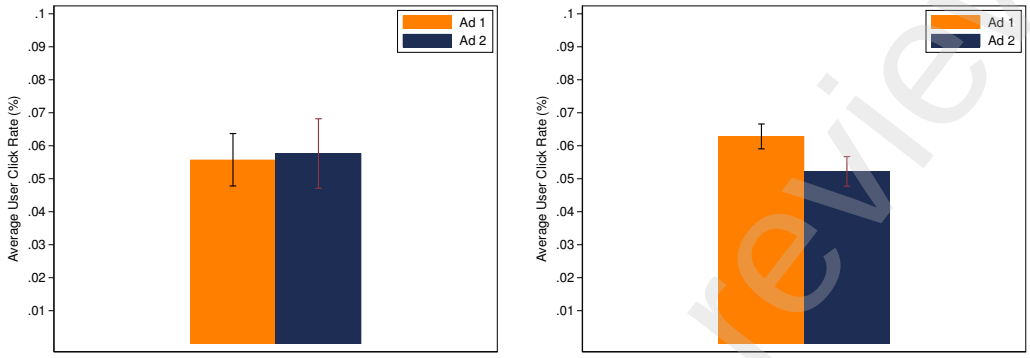
2. *Targeting Effect:* Richer user-level data allows the platform to better capture heterogeneity in preferences and show more personalized ads.

To illustrate the *Precision Effect*, consider Figure 3, which compares average click rates for two ads. Sub-Figure 3a shows estimates based on a 20% random subsample, while Sub-Figure 3b uses the full dataset. In the subsample, Ad 2 appears to have a slightly higher click rate, although the difference is not statistically significant. In contrast, using the full data reveals that Ad 1 has a significantly higher click rate. As a result, access to more data can change which ad is predicted to be optimal and improve allocation decisions. We refer to this mechanism as the *Precision Effect*.

The *Targeting Effect* is illustrated in Figure 4. Sub-Figure 4a shows that the two ads have similar average click rates across all users, with Ad 2 performing slightly better. However, Sub-Figure 4b conditions on whether users have visited political websites and reveals substantial heterogeneity. Users who visit political websites are significantly more likely to click on Ad 2, while the opposite holds for other users. This pattern demonstrates that richer user-level data allows the platform to better capture preference heterogeneity, show more personalized ads, and increase click rates.

Preprint not peer reviewed

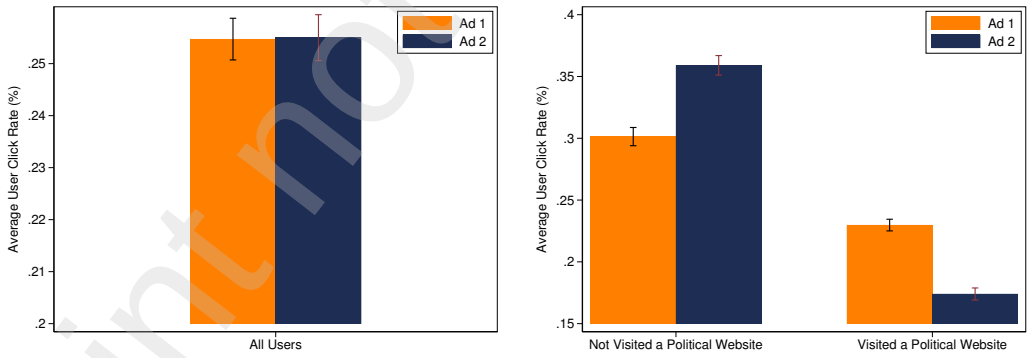
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076



(a) Based on a 20% Random Subsample (b) Based on All Observations

Figure 3: Average Click Rates for Two Ads and the Precision Effect

Notes: This figure depicts the raw average click rates and their 95% confidence intervals for ads ranked 10 and 19, respectively. Click rates are calculated as the number of clicks divided by the total number of impressions.



(a) Among all Users (b) Among Two Different User Groups

Figure 4: Average Click Rates for Two Ads and the Targeting Effect

Notes: This figure depicts the raw average click rates and their 95% confidence intervals for ads ranked 1 and 3, respectively. Click rates are calculated as the number of clicks divided by the total number of impressions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

4. Data

4.1. Data Source and Description

This study uses a novel dataset from the largest online advertising platform in Iran. The data consist of user-level observations of ad impressions, based on a 2.5% random sample of user devices that visited more than 2,000 partner websites between September 3 and 9, 2022. An impression is defined as a user's exposure to an ad. Users are tracked via cookies, allowing the platform to observe their browsing behavior, ad exposures, and click outcomes over time. The richness of the data enables detailed measurement of user-level interactions across websites and advertisements, which is central to estimating heterogeneous click behavior.

For each user device, the dataset captures a variety of details: the device type (mobile or desktop), all website visits (among the websites using the platform), the identity and subject matter of the websites visited, visit timestamps, the ads shown during each visit, the per-click bid amounts for those ads, and whether the user clicked on any shown ad.

4.1.1. User Device and Impression Characteristics

Table 1 summarizes user and impression characteristics. The dataset includes approximately 1.6 million devices, of which 91% are mobile, reflecting the dominance of mobile traffic in digital advertising. On average, users visit 38 pages and 4.2 distinct websites during the sample period. They are exposed to about 110 ads (40 distinct), indicating repeated exposure to the same ads.

There is substantial heterogeneity in user behavior. The standard deviations of pages visited and ads viewed are significantly larger than their means, suggesting wide variation in browsing intensity. Finally, the average user click-through rate is 0.089%, highlighting the rarity of user clicks in this setting.

4.1.2. Website Characteristics

The platform classifies partner websites into 13 content categories after aggregating smaller groups. Table 2 summarizes website characteristics by category. The distribution of websites and user activity varies substantially across categories. For example, music websites attract the largest number of users and have the highest average click-through rates, while categories such as political and childcare exhibit lower engagement.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Variable	Mean	SD
Number of Users	1,596,625	-
Number of Pages Visited	38.05	64.71
Number of Distinct Websites Visited	4.19	5.28
Number of Ads Seen	109.54	459.76
Number of Distinct Ads Seen	40.47	56.28
Share of Mobile Devices	0.91	0.28
User Click Rate (%)	0.089	0.264

Table 1: User Device and Impression Characteristics

Notes: Averages and standard deviations are calculated at the user level. The calculations are based on one week of data from September 3-9, 2022. Devices with more than 500 total visits (about 1% of the devices) are removed. SD stands for standard deviation.

These differences highlight variation in user attention and engagement across content types, which may influence ad allocation and click behavior.

4.1.3. Advertiser Characteristics

Table 3 presents summary statistics for the top 300 advertisers, which account for over 90% of total impressions. On average, each advertiser receives 0.7 million impressions and 482 clicks during the sample period. The average daily budget is $50.50, and the mean per-click bid is 1.96 cents.

There is considerable heterogeneity across advertisers. Clicks range from fewer than 10 to over 15,000, and daily budgets vary from near zero to almost $1,000. Similarly, per-click bids and cost-per-mille measures exhibit substantial dispersion, reflecting differences in advertiser valuations and strategies.

5. Model

To study this market outcome in different counterfactual scenarios, we model the behavior of different agents in the model:

5.1. User Click Rate Model

The key object of interest is the probability that a user clicks on a given ad. Let y_i^a denote an indicator for whether user impression i results in a click on ad a. The objective is to estimate the conditional click probability $\mathbb{E}(y^a|X_i)$, where X_i includes user and impression characteristics, for all a and i. We estimate click probabilities for the 100 most frequently shown ads in the dataset. The remaining ads, which have relatively few observations,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Subject	Websites	Visits	Users	Ads	Avg. CTR(%)
Political	204	27.20	592.36	84.50	0.047
Music	306	11.10	670.12	24.90	0.147
Health	127	8.96	526.37	31.40	0.067
Finance	45	7.65	345.53	21.50	0.079
Entertainment	139	5.34	395.49	15.50	0.096
Childcare	6	5.15	94.98	11.60	0.033
Sport	24	4.55	156.34	15.20	0.043
Other	923	3.09	288.22	8.26	0.086
Lifestyle	77	2.50	211.69	8.07	0.093
Movies	78	1.90	126.82	5.52	0.069
Tech	111	1.48	155.17	3.54	0.101
Education	61	0.89	110.01	2.04	0.067
Art	32	0.57	102.13	1.64	0.083
Total	2133	80.38	1,596.63	233.67	0.070

Table 2: Characteristics of Websites with Different Subjects

Notes: The calculations are based on one week of data from September 3-9, 2022. Websites, visits, users, and ads show the number of websites, total number of user visits, number of unique users, and the number of ads shown within each subject, respectively. Numbers of visits and ads are in millions and the number of users are in thousands. The number of ads is equal to the number of impressions. CTR stands for Click-Through Rates.

Variable	Mean	Min	Median	Max
Impressions (millions)	0.70	0.10	0.24	19.30
Clicks	481.66	8	149.5	15651
Daily Budget ($)	50.50	0.001	20	988
Per-Click Bid (Cents)	1.96	0.003	1.78	9.61
Avg. Click Rate (%)	0.06	0.006	0.057	0.21
Cost Per Million Impressions (Cents)	1230	0.25	934	6159

Table 3: Advertiser Characteristics Summary Statistics

Notes: The calculations are based on one week of data from September 3-9, 2022. All the statistics are unweighted. Impressions and clicks show the total number of times each advertiser's ads are shown and clicked on. Cost per million impressions is how much each advertiser pays on average to show its ads a million times. Monetary values are converted to US dollars using the average exchange rate in September 2022.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

are grouped into a single "fringe" category and treated as an additional ad (ranked 101). These estimates are used to determine which ads maximize expected revenue under different counterfactual scenarios.

To estimate heterogeneous click probabilities, we use the causal forest method of Athey et al. (2019). This approach allows for flexible estimation of conditional treatment effects in high-dimensional settings without imposing restrictive functional form assumptions.

This model relies on a linear probability specification:

$$y = g(X) + \tau(X).A + \epsilon, \;\; \mathbb{E}(\epsilon|X) = 0 \tag{2}$$

where $g(X)$ is the baseline click probability (the click rate for the ad considered as the base treatment), $\tau(X)$ is the conditional Average Treatment Effect (CATE) vector at X, and A is the treatment (ad) assignment variable vector such that

$$A_a = \begin{cases} 1 & \text{if ad } a \text{ is shown} \\ 0 & \text{otherwise} \end{cases}$$

Let $e(X) = \mathbb{E}(A|X)$ denote the propensity score, i.e., the probability that a given ad is shown conditional on observed characteristics. Because ad assignment depends on observable factors such as website and time, failing to account for propensity scores would bias estimates of click probabilities.

Following Robinson (1988), the endogeneity is resolved by residualizing y and A with respect to X. This yields a moment condition that identifies the heterogeneous treatment effect $\tau(X)$.

$$\mathbb{E}(\epsilon|X) = \mathbb{E}\left(y - m(X) - \tau(X).\left(A - e(X)\right)|X \right) = 0 \tag{3}$$

The sample equivalent of 3 is:

$$\hat{\tau}(x) = \text{argmin}_{\tau} \left(\sum_{i=1}^{n} \alpha_i(x) \left[y_i - \hat{m}^{(-i)}(X_i) - \tau.\left(A_i - \hat{e}^{(-i)}(X_i)\right) \right] \right)^2 \tag{4}$$

The *nuisance functions* $\hat{m}(X)$ and $\hat{e}(X)$ are estimated nonparametrically, and treatment effects are recovered via a local regression using causal forest weights. Also, the superscript $(-i)$ indicates that the nuisance functions are estimated using a sample that excludes observation i, implementing a leave-one-out (cross-fitting) procedure to avoid overfitting and ensure unbiased

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

estimation. Please refer to the appendix for more details about the model estimation.

Finally, the click probabilities can be estimated using the following equation:

$$\hat{\mathbb{E}}(y^a|x) = \hat{m}(x) + \hat{\tau}(x).\,(A - \hat{e}(x)) \tag{5}$$

There are two main identifying assumptions for the treatment effect estimates to be consistent (Athey and Imbens (2016); Wager and Athey (2018)):

1. *Unconfoundedness* (or randomization conditional on covariates): The treatment allocation must be independent of potential outcomes conditional on observables:
$$A_i \perp Y_i(a)|X_i \qquad \forall a \in A$$
2. *Overlap*: Every available treatment must have a positive probability of being received for all $x \in \boldsymbol{X}$. So, there must be some $\epsilon > 0$ such that:
$$\epsilon < \mathbb{P}[A = a|X = x] < 1 - \epsilon$$

The first assumption states that treatment assignment depends only on observables. This condition is satisfied under the maintained assumption that ad assignment depends only on observed characteristics, such as website and time, as discussed in Section 3.1.1.

The overlap assumption requires that all ads have a possible propensity of being shown to all users. This condition is also satisfied based on equation 1. Some empirical evidence supporting this feature is provided in the appendix.

5.1.1. Inputs to the Click Probability Model

This section describes the user and impression characteristics used to estimate conditional click probabilities. We distinguish between two types of variables: (i) historical user characteristics and (ii) current impression characteristics. This distinction is important for two reasons. First, only current impression characteristics affect ad assignment and thus propensity scores, and second, the platforms always observe the characteristics of current user impressions regardless of the market structure. So they are not affected by a merger or breakup.

We briefly discuss these characteristics here. More detailed discussion is provided in the appendix.

Historical User Characteristics : These include browsing history, past ad exposure, prior clicks, and overall user activity.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

- **Browsing History**: Indicators for past visits to the 13 website categories to capture user interests and demographics.
- **Previous Exposures**: The number of prior exposures to an ad accounts for diminishing returns to repeated advertising.
- **Previous Clicks on the Ad**: Prior clicks indicate user interest and increase the likelihood of future clicks.
- **User Activity Level**: More active users tend to be less responsive to ads.

Current Impression Characteristics These include features observed at the time of the impression and affect both ad allocation and click probabilities.

- **Website Category**: 13 indicator variables representing the subject matter of the current website.
- **Website Rank within Category**: Captures within-category variation in engagement.
- **Number of Ad Slots**: More ads dilute user attention and reduce click probability.
- **Time of Impression**: Captures temporal variation in user behavior and ad allocation.
- **User Device Type**: Controls for differences between mobile and desktop users.

5.2. Platform(s) Model

The platform's objective is to maximize expected revenue from each user impression by selecting which ad to display. Because advertisers pay on a per-click basis, expected revenue depends on both the likelihood of a click and the advertiser's per-click payment.

The expected revenue from showing ad a to impression i is given by:

$$\text{Expected Revenue}(a, i) = \text{Per-click Price} \times \mathbb{E}(y^a | X_i)$$

Under **uniform pricing**, all advertisers face the same per-click price. In this case, maximizing expected revenue is equivalent to selecting the ad with the highest predicted click probability:

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

$$a_i = \text{argmax}_{a \in \mathcal{A}} \mathbb{E}(y^a | X_i)$$

Under **non-uniform pricing**, advertisers have heterogeneous per-click valuations. We assume the platforms know advertisers' valuations and can charge them accordingly. The platform then selects the ad that maximizes expected revenue:

$$a_i = \text{argmax}_{a \in \mathcal{A}} \{v_a \times \mathbb{E}(y^a | X_i)\}$$

This specification allows the platform to account for both user preferences and advertiser valuations when allocating ads.

Also, when multiple ad slots are available on a given page (e.g., n slots), we assume that the platform displays the top n ads ranked by expected revenue. The click probabilities $\mathbb{E}(y^a|X_i)$ are estimated using the causal forest model described in Section 5.1, allowing for flexible heterogeneity across users and ads.

The analysis focuses on static optimization. Incorporating dynamic learning would require additional assumptions about platform behavior and advertiser responses that are not possible given the data limitations.

This framework maps estimated user-level click probabilities into platform-level allocation decisions, allowing for simulation of counterfactual market outcomes under different data environments.

5.3. Advertisers Model

Advertisers value user clicks, which represent potential conversions. We assume that advertisers differ in their per-click valuations v_a , reflecting differences in profitability, customer lifetime value, and marketing objectives. Under uniform pricing, all advertisers face the same per-click price. We assume that this price is below all advertisers' valuations, ensuring full participation. Under non-uniform pricing, advertiser valuations are required to compute expected revenue. We recover these valuations under additional assumptions, described in the appendix.

6. Results

6.1. Click Rate Estimates

Figure 5 presents the cumulative distribution of estimated click probabilities for the top ads. Click probabilities are highly skewed, with most impressions

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

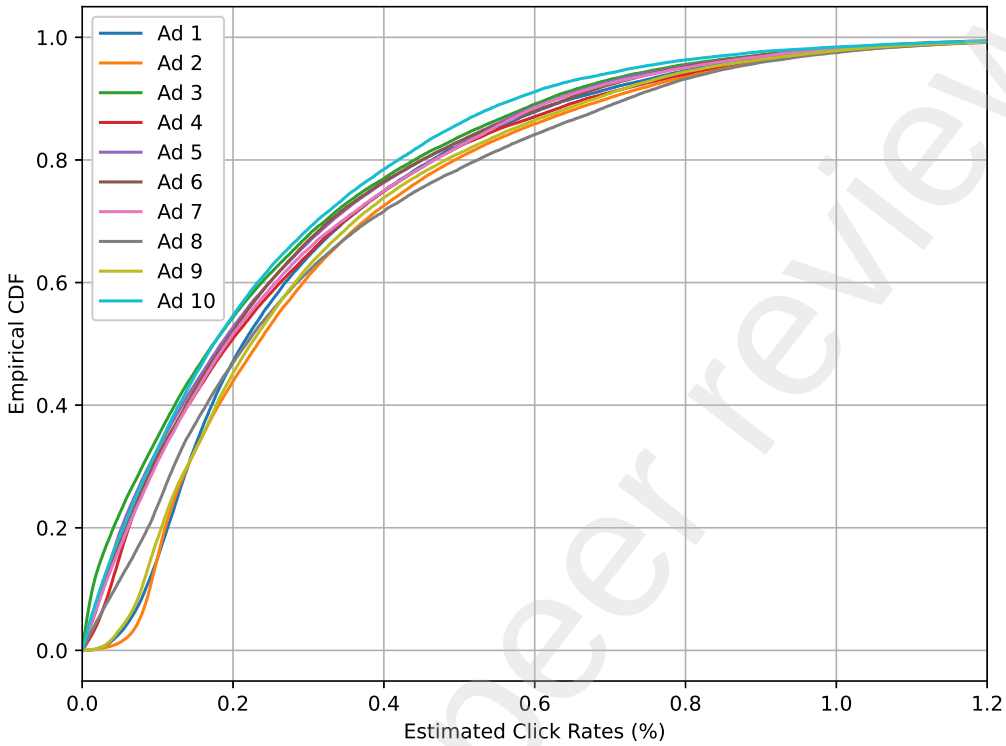


Figure 5: Empirical CDF of the Estimated Click Rates for Top 10 Ads

Notes: This figure shows the empirical cumulative distribution functions (CDFs) of estimated click-through rates (CTR) for the top 10 ads. Each curve represents one ad and is constructed using predicted click probabilities at the user-impression level. The steep increase in the CDFs at low values highlights the rarity of clicks, while variation across curves reflects differences in ad performance and substantial heterogeneity in user responses. Ads are ranked by their total number of impressions.

having low likelihood of a click but a non-trivial share exhibiting much higher probabilities. This dispersion reflects substantial heterogeneity in user preferences across ads.

This heterogeneity is economically important: if platforms can identify high-probability matches at the user level, they can significantly improve overall click-through rates. The results therefore highlight the potential gains from personalized ad allocation relative to uniform or aggregate targeting strategies.

To further illustrate preference heterogeneity, Figure 6 compares click probabilities across pairs of ads. For each pair, the figure reports the share of impressions for which one ad has a higher predicted click probability than the other. Values close to 0.5 indicate substantial heterogeneity, as neither

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

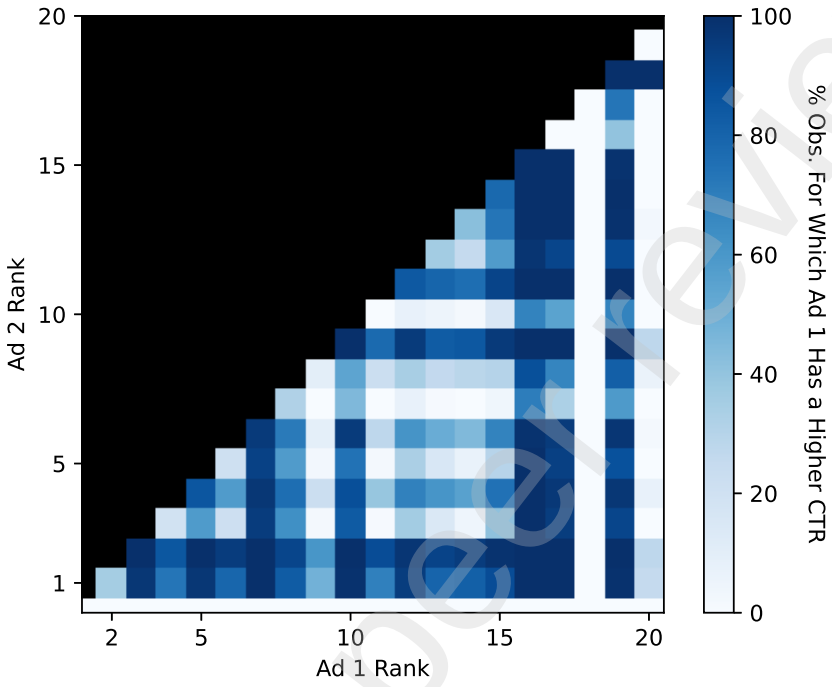


Figure 6: Share of Observations for Which Ad 1 has a higher CTR than Ad 2

Notes: Share of observations for which Ad 1 (on the horizontal) has a higher predicted click-through rate than Ad 2 on the vertical), for each pair of ads. Ads are ranked by their total number of impressions. The dispersion around 50% for many ad pairs indicates significant heterogeneity in user-level preferences, implying that optimal ad allocation varies across users.

ad consistently dominates.

The results show that for most ad pairs, neither ad uniformly outperforms the other across users. This reinforces the importance of personalization: the optimal ad choice varies across users, and targeting can significantly improve match quality.

6.2. User Characteristics Affecting the Click Rate Heterogeneities

To understand the sources of heterogeneity in click behavior, we examine the contribution of different user and impression characteristics using *Shapley values* and *feature importance* measures.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

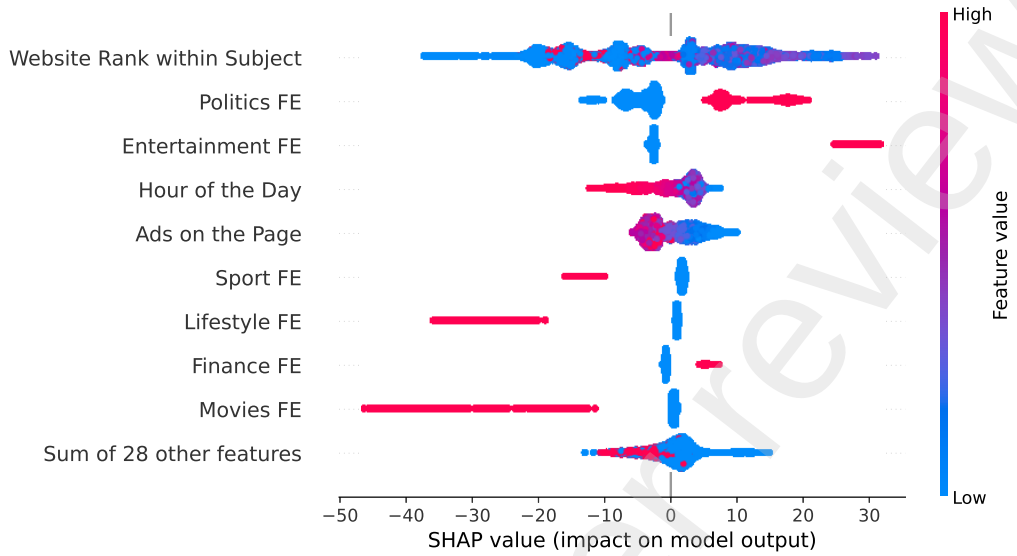


Figure 7: Estimated Shapley Values for the Treatment Effect between Ads Ranked 1, 2

Notes: Shapley value measures the average marginal impact of each user feature on the outcome variable (the treatment effect here). The treatment effect is the expected probability of clicking on ad 1 minus the expected probability of clicking on ad 2. Ads are ranked by their total number of impressions. The Shapley values are normalized by the average click-through rate of the second ad and can be interpreted as the percentage change in click rate attributable to each feature. FE denotes fixed effects (for the current website subject).

6.2.1. Shapley Values

We first use Shapley values to measure the contribution of each feature to predicted treatment effects. Intuitively, Shapley values capture the marginal importance of a variable by averaging its contribution across different model specifications.

Figure 7 presents Shapley values for the treatment effect (difference in click probabilities) between two ads. The results show that website characteristics and exposure-related variables play a central role in shaping treatment effects. In particular, the rank of the website within its category has a strong and non-linear impact, while certain content categories systematically increase or decrease the relative effectiveness of ads.

Content categories such as political, entertainment, and finance are associated with higher treatment effects (higher click rates for ad 1), while others such as lifestyle and music are associated with lower relative performance (higher chances of clicking on ad 2). Time of day and the number of ads on the page also affect click probabilities, although their contributions are smaller.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

6.3. Feature Importances

To complement the Shapley analysis, we compute feature importance measures that summarize the contribution of each variable to overall heterogeneity in treatment effects.

Figure 8 shows that a small number of variables account for most of the variation in treatment effects. In particular, ad exposure variables, the number of ads on the page, and website rank (which can be interpreted as a website fixed effect) jointly account for the majority of heterogeneity. These findings highlight the importance of both user history and contextual features in determining ad effectiveness. This suggests that effective targeting relies on combining behavioral data with contextual information.

7. Counterfactual Simulation

7.1. User and Platform Behavior Simulation

To evaluate the effects of data on market outcomes, we simulate user behavior and platform allocation under alternative data environments. Because user responses depend on past ad exposure and platform decisions depend on predicted click probabilities, the simulation captures the interaction between user behavior and platform optimization.

The simulation is initialized using observed user behavior in the first five days of the data. These initial conditions determine users' browsing histories and past ad exposures, which are key inputs into click probability predictions. The last two days of the data are used to estimate click rates and conduct counterfactual simulations. The simulations are conducted on a 20% random subsample of users who visited partner websites on September 8 and 9, 2022, comprising 160,394 users, 965,265 total visits, and 5,249,146 impressions.

7.2. Simulation Procedure

For each user impression, the simulation proceeds as follows:

1. Predicted click probabilities are computed using the estimated model given the data available to each platform.
2. Each platform selects the ads that maximize expected revenue under the specified pricing rule.
3. A click outcome is realized based on the actual (full-data) probability.
4. User history is updated and used for subsequent impressions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

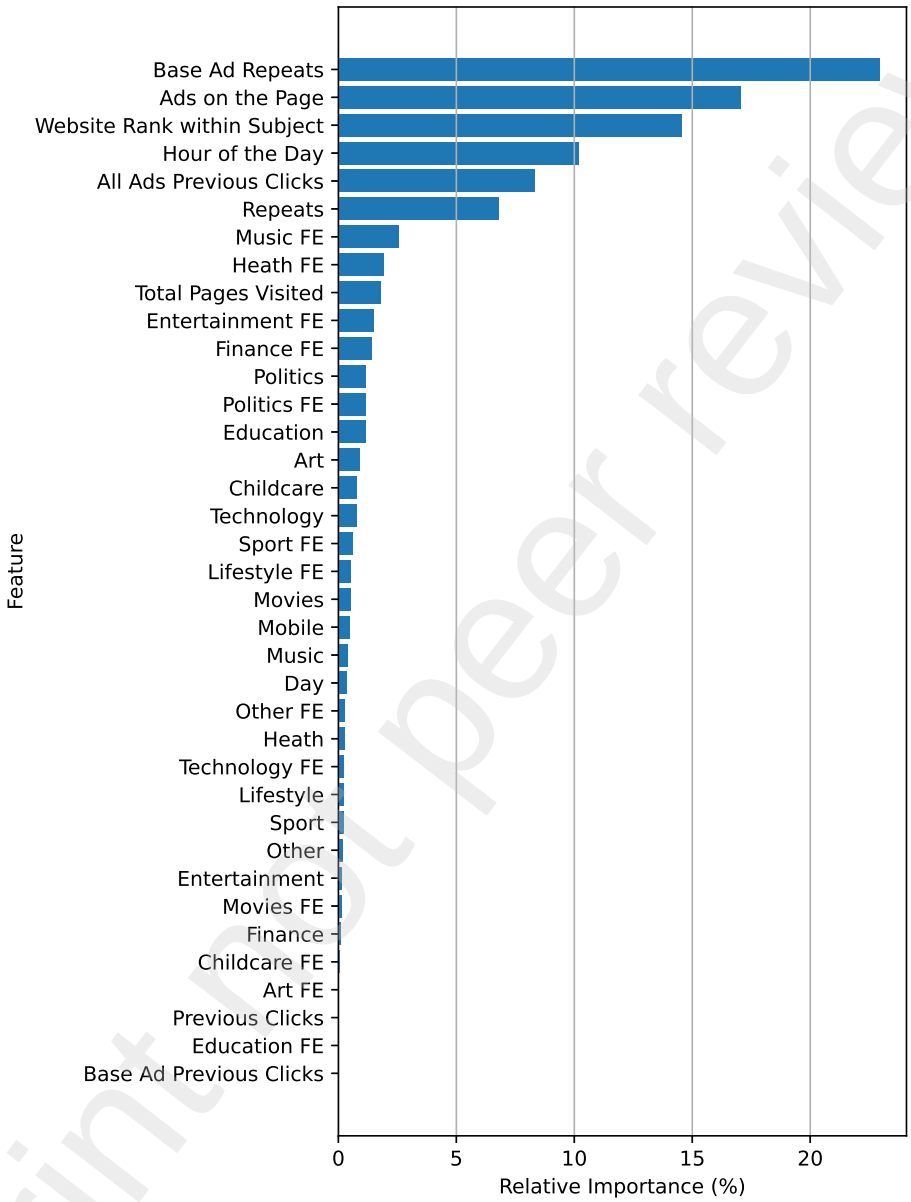


Figure 8: Average Relative Importances of User Characteristics

Notes: Relative importances are measured by the average heterogeneity each characteristic creates across all the estimated CATEs. The values are normalized, so they sum up to 100%. FE denotes fixed effects for the current website subject. Subject-level variables without "FE" capture users' historical visits to each subject group. Base Ad Repeats denotes the number of times the base (treatment) ad is shown. Repeats denotes the number of times the comparison ad is shown. Day denotes day fixed effects. All Ads Previous Clicks denotes the number of prior clicks across all ads. Hour of the Day is a continuous variable ranging from 0 to 23 capturing the time of the impression. More details on the calculation are provided in the appendix.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

7.3. Counterfactuals

We consider several counterfactual scenarios corresponding to different market structures and data availability.

In the *monopoly* scenario, the platform observes the full set of user interactions across all websites. This represents the case in which all available data are pooled, allowing for the most accurate prediction of user behavior. In the *duopoly* scenarios, user data are partitioned across two platforms, so that each platform observes only a subset of user interactions. This limits their ability to learn user preferences and target ads effectively. We consider several alternative partitions that restrict information along different dimensions:

1. **Random Split by Website:** Websites are randomly divided between the two platforms, and each platform serves ads only on its assigned websites. As a result, each platform observes user behavior only on a subset of websites.
2. **Random Split by Website Subject:** Websites are divided based on their content categories (e.g., sports, politics, entertainment). Each platform observes user behavior only within its assigned set of categories, limiting its ability to track users across different types of content.
3. **Random Split by User:** Users are randomly assigned to one platform. Each platform observes the full browsing behavior of its assigned users across all websites, but has no information about users assigned to the other platform.

In all counterfactuals, allocation is kept balanced to ensure comparable numbers of impressions across platforms. Additional details on the construction of these scenarios are provided in the appendix.

These three splits differ in the dimensions along which platform information is restricted: the set of websites, the type of content, or the group of users. Table 4 summarizes the data available in each scenario, including whether platforms observe the full set of ad impressions, complete user histories, and cross-category browsing behavior.

These distinctions determine how data fragmentation affects prediction accuracy, targeting effectiveness, and overall market efficiency.

7.4. Counterfactual Results under Uniform Pricing

Table 5 reports the counterfactual average click-through rates (CTR) under uniform pricing across different data environments. The results show that access to additional data substantially improves match quality.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Description	Full Observations	Full Data on Each User	Full Data on Available Subjects
Monopoly	✓	✓	✓
Random Split by Website	x	x	x
Random Split by Website Subject	x	x	✓
Random Split by User	x	✓	✓

Table 4: Summary of the Characteristics of Data Splits in Counterfactuals

Under the monopoly scenario, where the platform observes the full set of user interactions, the average CTR reaches 0.372%, more than double the status quo level of 0.170%. This corresponds to a 118.8% increase in match quality, highlighting the large efficiency gains from pooling user data.

In contrast, data fragmentation reduces performance across all duopoly scenarios. When data are split across platforms, average CTR declines to 0.310%–0.324%, depending on how information is partitioned. This implies that mergers increase CTR by 14.1%-20% relative to these fragmented environments.

The decline in click-through rates arises from selecting suboptimal ads under fragmented data scenarios. To quantify how often the optimal ad is selected, we define a similarity measure as the share of impressions for which the ad chosen under a given counterfactual matches the ad selected under the monopoly benchmark. The value of this measure varies across scenarios. Splitting data by user leads to the lowest similarity with the monopoly benchmark (37.7%), indicating a substantial loss of information about individual user behavior. Splitting by website yields a slightly higher similarity (46.7%), while splitting by website subject performs best among duopolies (52.6%), suggesting that retaining information within content categories partially preserves targeting ability.

These results highlight the importance of data availability for platform performance, as all forms of fragmentation lead to meaningful reductions in match quality relative to the monopoly benchmark. At the same time, differences across the duopoly scenarios are relatively modest, suggesting that the specific way in which data is fragmented plays a limited role compared to the overall loss of data.

7.5. Counterfactual Results under Non-Uniform Pricing

This section examines counterfactual outcomes under non-uniform pricing, where advertisers have heterogeneous per-click valuations. In this setting, the platform selects ads to maximize expected revenue, taking into account both

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

No.	Description	Average CTR (%)	Similarity with Monopoly (%)	CTR Gains After Merger (%)
1	Monopoly	0.372	100	–
2	Random Split by Website	0.310	46.7	20
3	Random Split by Website Subject	0.324	52.6	14.1
4	Random Split by User	0.311	37.7	19.6
	Status Quo	0.170	31.8	118.8

Table 5: Counterfactual Average Click Rates (with Uniform Pricing Assumption)
Notes: CTR stands for click-through rate. Similarity with monopoly measures the share of observations for which the optimal ad under monopoly is shown. CTR gains after merger are the gains when switching from each counterfactual to monopoly. The simulations are conducted on a 20% random subsample of users who visited partner websites from September 8-9, 2022, consisting of 160,394 users, 965,265 total visits, and 5,249,146 impressions. The status quo average click-through rate is higher than the overall average described in the data section, as the simulation sample is drawn from the final two days of the data, which fall on a weekend.

click probabilities and advertiser valuations. The results are qualitatively similar to those under uniform pricing. Access to additional data continues to substantially improve match quality, while data fragmentation reduces performance across all duopoly scenarios.

Under the monopoly scenario, the average click-through rate (CTR) reaches 0.354%, compared to 0.170% in the status quo, corresponding to a 108.2% increase in match quality. In contrast, when data are fragmented across platforms, average CTR declines to between 0.282% and 0.295%, depending on how information is partitioned. This implies that mergers increase CTR by 20% to 25.5% relative to these duopoly environments.

As in the uniform pricing case, differences across duopoly scenarios are modest. Splitting data by website yields the lowest similarity to the monopoly benchmark (31.64%), followed by splitting by user (33.37%) and by website subject (40.09%). This pattern suggests that restricting access to cross-context information reduces targeting performance, although no single form of fragmentation dominates.

Relative to uniform pricing, click-through rates are lower, as the platform's objective shifts from maximizing click probabilities to maximizing expected revenue. However, the main conclusions remain unchanged: access to larger datasets significantly improves match quality, and data fragmentation leads to meaningful efficiency losses across all scenarios.

7.6. Decomposing the Sources of Match Quality Gains

The previous sections show that access to additional data substantially improves match quality. This section decomposes these gains into two distinct

Preprint, not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

No.	Description	Average CTR (%)	Similarity with Monopoly (%)	CTR Gains After Merger (%)
1	Monopoly	0.354	100	–
2	Random Split by Website	0.282	31.6	25.5
3	Random Split by Website Subject	0.295	49.1	20
4	Random Split by User	0.286	33.4	23.8
	Status Quo	0.170	31.8	108.2

Table 6: Counterfactual Average Click Rates (with Non-Uniform Pricing Assumption)
Notes: CTR stands for click-through rate. Similarity with monopoly measures the share of observations for which the optimal ad under monopoly is shown. CTR gains after merger are the gains when switching from each counterfactual to monopoly. The simulations are conducted on a 20% random subsample of users who visited partner websites from September 8-9, 2022, consisting of 160,394 users, 965,265 total visits, and 5,249,146 impressions. The status quo average click-through rate is higher than the overall average described in the data section, as the simulation sample is drawn from the final two days of the data, which fall on a weekend.

mechanisms: improvements in prediction accuracy and improvements in the ability to exploit user heterogeneity. As discussed in Section 3.3, a data-driven merger can improve match quality through two channels. First, a *Precision Effect*: access to more observations allows the platform to estimate click probabilities more accurately, reducing errors in selecting the optimal ad. Second, a *Targeting Effect*: access to richer user-level data enables the platform to better capture heterogeneity in user preferences and tailor ads more effectively.

These two mechanisms capture different roles of data. The Precision Effect reflects gains from increasing the sample size used for prediction, while the Targeting Effect reflects gains from expanding the set of observable user characteristics.

To separate these effects, we construct intermediate counterfactual scenarios that vary the amount of data used for prediction while holding the set of observable user characteristics fixed, and vice versa. This approach allows us to isolate the contribution of each mechanism to overall improvements in match quality.

In particular, we compare outcomes across scenarios in which platforms have access to (i) limited observations with full observables, (ii) full observations with limited observables, and (iii) both full observations and full observables. These comparisons isolate the marginal contribution of each channel.

Table 7 reports average click-through rates (CTR) under intermediate counterfactual scenarios that isolate the contributions of additional observations and additional observables (under uniform pricing assumption).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

The monopoly benchmark, which combines full observations and full observables, achieves a CTR of 0.372%. In contrast, when both observations and observables are limited (baseline duopoly scenarios), CTR declines substantially to 0.310% under website splits and 0.324% under website-subject splits.

To isolate the role of additional observables (Targeting Effect), we compare scenarios with limited observations but full observables. In these cases, CTR increases modestly from 0.310% to 0.322% under website splits and from 0.324% to 0.332% under website-subject splits. These improvements indicate that expanding the set of observable user characteristics leads to relatively small gains in match quality.

In contrast, to isolate the role of additional observations (Precision Effect), we compare scenarios with full observations but limited observables. Here, CTR increases more substantially, from 0.310% to 0.360% under website splits and from 0.324% to 0.361% under website-subject splits. These gains are much closer to the monopoly benchmark, suggesting that increasing the volume of data used for prediction has a much larger impact on match quality.

These results show that the majority of the gains from additional data are driven by improvements in prediction accuracy rather than by expanded observability. Across scenarios, increases in observables account for only a small fraction of the total improvement, while increases in observations explain most of the gap between fragmented and monopoly outcomes.

No.	Description	Full Observations	Full Observables	Average CTR (%)
1	Monopoly	✓	✓	0.372
2	Random Split by Website	x	x	0.310
2'	Random Split by Website	x	✓	0.322
2"	Random Split by Website	✓	x	0.360
3	Random Split by Website Subject	x	x	0.324
3'	Random Split by Website Subject	x	✓	0.332
3"	Random Split by Website Subject	✓	x	0.361

Table 7: Average Click Rates Under Different Counterfactuals (with Uniform Pricing Assumption)

Notes: CTR stands for click-through rate. The simulations are conducted on a 20% random subsample of users who visited partner websites from September 8-9, 2022, consisting of 160,394 users, 965,265 total visits, and 5,249,146 impressions.

These findings should be interpreted in the context of the specific platform and dataset used in this study. In particular, the number of observations available in the data is substantially smaller than in large-scale real-world

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

platforms, which may affect the relative importance of the precision effect. The number of ads in the dataset is also smaller than in typical online advertising markets, potentially reducing the scope for differentiation across ads. Moreover, the limited time horizon of the data restricts the ability to fully capture users' preferences and browsing histories, potentially attenuating the role of targeting. As a result, the relative contributions of precision and targeting may differ in settings with larger datasets, a richer set of ads, and more extensive user histories.

8. Conclusion

This paper studies how access to data affects match quality in online advertising markets, with a particular focus on how mergers between platforms with complementary datasets—or the breakup of a monopolistic platform—shape market outcomes. Using detailed data from Iran's largest online advertising platform and leveraging a conditionally random ad allocation mechanism, we estimate heterogeneous click probabilities and simulate counterfactual market structures.

The results show that data plays a central role in improving match quality. Data-driven mergers increase click-through rates by 14–25%, while breaking up a monopolistic platform leads to substantial efficiency losses. A key contribution of the paper is to decompose these gains into two mechanisms. The findings indicate that most of the improvement is driven by increased data volume (*Precision Effect*), while expanding the set of observable user characteristics (*Targeting Effect*) plays a more limited role. This pattern is particularly important in environments such as display advertising, where click rates are extremely low and accurate estimation requires large numbers of observations. Conversion rates are typically even orders of magnitude lower than click rates, implying that the importance of large datasets is even greater when the objective is to maximize conversions.

These findings have important implications for market design and policy. On the one hand, increased data concentration can generate substantial efficiency gains by improving ad allocation and match quality. On the other hand, greater market power may enable platforms to raise prices or engage in price discrimination, potentially harming advertisers. The estimated efficiency gains suggest that mergers may improve advertiser welfare if price increases remain moderate, while platform breakups could be welfare-enhancing only if they lead to sufficiently large reductions in prices. Policymakers should

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

therefore weigh the trade-off between efficiency and market power carefully when evaluating antitrust interventions in data-driven markets.

At the same time, the results should be interpreted with caution. The analysis is based on data from a single platform operating in a relatively isolated market, and the findings may not generalize directly to larger global platforms such as Google or Meta. In larger markets, diminishing returns to data or differences in the scale and diversity of user activity may alter the relative importance of data volume and observability. Understanding how these forces operate at scale remains an important direction for future research.

This study also has several limitations. First, click-through rates are used as the primary measure of match quality, which may not fully capture user engagement or advertiser value. Future work could incorporate outcomes such as conversions or long-term customer behavior. Second, the short data horizon limits the ability to construct richer measures of user history, and longer panels could provide deeper insight into dynamic learning and targeting. Third, the analysis assumes a fixed number of ad slots and treats the set of ads as exogenous, while in practice both may respond to changes in market structure. In addition, the model assumes a fixed set of ads, whereas in practice improvements in targeting may encourage entry of more specialized or "niche" ads tailored to specific users. Finally, the model focuses on static revenue maximization and abstracts from the trade-off between experimentation and optimization, which is central to platform learning and could affect the value of data.

Several directions for future research emerge from this analysis. First, an important extension would be to study the optimal balance between experimentation and optimization in platform design. Platforms face a trade-off between exploring new ad-user matches to improve learning and exploiting existing information to maximize short-run revenue, and this trade-off may critically shape the value of data. Second, understanding the optimal design of antitrust interventions in data-driven markets remains an open question. In particular, determining when and how to implement platform breakups—taking into account both efficiency losses from data fragmentation and potential gains from increased competition—would be a valuable direction for policy-oriented research. Finally, incorporating dynamic considerations into the analysis could provide deeper insights into how market structure affects long-run outcomes. Changes in data availability may influence not only current match quality but also future learning, entry of advertisers, and the evolution

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

of platform competition over time.

Declaration of AI Use

The authors used ChatGPT (OpenAI) to assist with language editing and improving the clarity of the manuscript. The authors take full responsibility for the content of the manuscript.

References

Ali, S.N., Lewis, G., Vasserman, S., 2023. Consumer Control and Privacy Policies. AEA Papers and Proceedings 113, 204–209.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. The Annals of Statistics 47, 1148–1178. Publisher: Institute of Mathematical Statistics.

Bajari, P., Chernozhukov, V., Hortacsu, A., Suzuki, J., 2019. The Impact of Big Data on Firm Performance: An Empirical Investigation. AEA Papers and Proceedings 109, 33–37.

Brynjolfsson, E., Mitchell, T., 2017. What can machine learning do? workforce implications. Science 358, 1530–1534.

Calvano, E., Calzolari, G., Denicolo, V., Pastorello, S., 2023. Artificial intelligence, algorithmic recommendations and competition. Algorithmic Recommendations and Competition (May 14, 2023) .

Chen, Z., Choe, C., Cong, J., Matsushima, N., 2022. Data-driven mergers and personalization. The RAND Journal of Economics 53, 3–31.

Cremer, J., Crawford, G.S., Dinielli, D., Fletcher, A., Heidhues, P., Schnitzer, M., Morton, F.M.S., 2023. Fairness and Contestability in the Digital Markets Act. Yale Journal on Regulation 40, 973–1012.

Dubé, J.P., Misra, S., 2017. Scalable price targeting. w23775, National Bureau of Economic Research Cambridge, MA.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Farboodi, M., Veldkamp, L., 2021. A Model of the Data Economy.

Goldfarb, A., Tucker, C., 2019. Digital Economics. Journal of Economic Literature 57, 3–43.

Hagiu, A., Wright, J., 2023. Data-enabled learning, network effects, and competitive advantage. The RAND Journal of Economics 54, 638–667.

Jones, C.I., Tonetti, C., 2020. Nonrivalry and the Economics of Data. American Economic Review 110, 2819–2858.

Mirrokni, V., Muthukrishnan, S., Nadav, U., 2010. Quasi-proportional mechanisms: Prior-free revenue maximization, in: Latin American Symposium on Theoretical Informatics, Springer. pp. 565–576.

Prat, A., Valletti, T., 2022. Attention Oligopoly. American Economic Journal: Microeconomics 14, 530–557.

Robinson, P.M., 1988. Root-n-consistent semiparametric regression. Econometrica: Journal of the Econometric Society , 931–954.

Shiller, B., Waldfogel, J., 2009. Music for a song: An empirical look at uniform song pricing and its alternatives. Technical Report. National Bureau of Economic Research.

Shiller, B.R., et al., 2013. First degree price discrimination using big data. Brandeis Univ., Department of Economics.

Tambe, P., Hitt, L., Rock, D., Brynjolfsson, E., 2020. Digital Capital and Superstar Firms.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

Wernerfelt, N., Tuchman, A., Shapiro, B., Moakler, R., 2024. Estimating the Value of Offsite Tracking Data to Advertisers: Evidence from Meta. Technical Report. National Bureau of Economic Research.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Online Appendix for "Should We Share Our Cookies? The Effect of Data on Match Quality in Online Advertising"

This appendix provides additional details on variable construction, estimation procedures, and robustness analyses supporting the main results.

A1. More Details on User Model Specification

The estimation is implemented in Python using the `econml` package, which provides tools for causal machine learning methods such as causal forests. This model relies on a linear probability specification:

$$y = g(X) + \tau(X).A + \epsilon, \;\; \mathbb{E}(\epsilon|X) = 0 \tag{A.1}$$

where $y \in \{0, 1\}$ denotes the outcome of interest, in this case, whether a user clicks on an ad, $X \in \mathbb{R}^{\mathrm{p}}$ is the set of user impression and website characteristics, $g(X)$ is the conditional average click rate for the base ad (base treatment), $\tau(X)$ represents the conditional Average Treatment Effect (CATE) vector at X, and A is the treatment (ad) assignment variable vector such that

$$A_a = \begin{cases} 1 & \text{if ad } a \text{ is shown} \\ 0 & \text{otherwise} \end{cases} \tag{A.2}$$

The propensity score vector is defined as the probability of receiving treatment a conditional on X, and it is denoted by $e(X)$. Element a of this vector indicates the probability of a user with characteristics X receiving treatment (being shown ad) a.

$$e(X) \triangleq \left[\Pr(A_1 = 1) \;\; ... \;\; \Pr(A_{n_A - 1} = 1)|X \;\right] = E(A|X) \tag{A.3}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Endogeneity arises because the observed treatment assignment is correlated with impression characteristics. Note that since the platform is currently not using targeting based on user's previous browsing and ad exposure characteristics, $e(.)$ is only a function of current user impression characteristics. We modify the model to make sure this holds in the estimation process. To control for endogeneity of ad allocation, the following method is used:

First, take an expectation from A.1 conditional on X to get:

$$\mathbb{E}(y|X) = \tau(X).\mathbb{E}(A|X) + g(X) \tag{A.4}$$

By defining $m(X) \triangleq E(y|X)$ as the expected click rate given X (not conditional on the treatment received) and by using A.3, equation A.4 can be written as:

$$m(X) = \tau(X).e(X) + g(X) \tag{A.5}$$

By subtracting A.5 from A.1 we will have:

$$y - m(x) = \tau(X).\left(A - e(X)\right) + \epsilon \tag{A.6}$$

And the identifying moment condition would be:

$$\mathbb{E}(\epsilon|X) = \mathbb{E}\left(y - m(X) - \tau(X).\left(A - e(X)\right)|X\right) = 0 \tag{A.7}$$

The sample equivalent of A.7 is:

$$\hat{\tau}(x) = \text{argmin}_{\tau}\left(\sum_{i=1}^{n} \alpha_i(x)\left[y_i - \hat{m}^{(-i)}(X_i) - \tau.\left(A_i - \hat{e}^{(-i)}(X_i)\right)\right]\right)^2 \tag{A.8}$$

Equation A.8 can be considered a local residual on residual regression. In other words, to estimate the treatment effect, which is the causal effect of A on y when both A and y are functions of X, the residuals of y after projecting on X are projected on the residuals on A after being projected on X. However, $\hat{m}(x)$ and $\hat{e}(x)$, which are expectations of Y and A conditional on $X = x$ (also noted as nuisance functions), are estimated nonparametrically in the first stage of the estimation. We use random forest regressor and random forest classifier models to flexibly estimate $\hat{m}(.)$ and $\hat{e}(.)$ functions, respectively.

Note that $(-i)$ superscripts in equation A.8 denote *"out of bag"* (OOB) predictions, meaning that observation i is not used to compute $\hat{m}^{(-i)}(X_i)$

Preprint, not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

and $\hat{e}^{(-i)}(X_i)$. Out-of-bag (OOB) prediction is used to further reduce overfitting and enable valid inference. Each causal forest is an ensemble of many trees built on bootstrap samples. For any given observation, its predicted treatment effect is computed by averaging the estimates across the subset of trees in which that observation was not included in the bootstrap sample. This prevents "self-prediction" and results in out-of-sample treatment effect estimates for each observation. OOB predictions are particularly useful for calculating valid confidence intervals and variance estimates for the CATE, as they provide a natural form of internal cross-validation.

Complementing out-of-bag prediction, honest estimation is a core structural property of causal forests that further improves the validity of treatment effect estimates. An honest tree separates the data used to determine the structure of the tree, that is, which covariates to split on and where, from the data used to estimate treatment effects within the leaves. Specifically, each tree randomly splits its training sample into two disjoint subsets: one subset is used exclusively for learning the tree's partitioning of the covariate space, and the other is used to estimate average treatment effects in each resulting leaf. This design prevents the model from tailoring splits to maximize fit on the same data used for effect estimation, thereby reducing overfitting and selection bias. Honesty has been shown to be necessary for valid confidence intervals and inference in forest-based treatment effect estimation. In combination with out-of-bag prediction, honesty allows causal forests to satisfy the orthogonality conditions needed for valid estimation in the presence of high-dimensional covariates and complex treatment effect heterogeneity.

In addition, importance weights are used to find the local treatment effect. $\alpha_i(x)$ is the relative importance weight of observation i for estimating the CATE at $X = x$, measuring the similarity of observation i and point x. These weights are determined using a causal forest algorithm in the second stage of the estimation. In this algorithm, a certain number of trees are trained based on random subsamples of the data. For each tree, the training data is split into two disjoint subsets: one used to build the tree structure (i.e., to choose the splits), and the other used for estimation within the leaves. Once the forest is constructed, $\alpha_i(x)$ is defined as the fraction of trees in which observation i and target point x fall in the same leaf (in the estimation sample). More precisely, if there are T trees and $L_t(x)$ denotes

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

the leaf containing x in tree t, then:

$$\alpha_i(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{i \in L_t(x)\}$$

This formula assigns more weight to observations that frequently co-occur in the same leaf as x, thereby capturing local similarity in the covariate space.

A1.1. Estimation Procedure

According to equation A.8, the treatment effect is estimated using the following moment condition:

$$\hat{\tau}(x) = \arg\min_{\tau} \left(\sum_{i=1}^{n} \alpha_i(x) \left[y_i - \hat{m}^{(-i)}(X_i) - \tau \cdot \left(A_i - \hat{e}^{(-i)}(X_i) \right) \right] \right)^2$$

Here, $\hat{m}^{(-i)}(X_i)$ is the predicted outcome under no treatment, estimated without using observation i; $\hat{e}^{(-i)}(X_i)$ is the predicted propensity score, also estimated without using i; and $\alpha_i(x)$ is the importance weight measuring how similar observation i is to the target point x.

A key strength of this estimation strategy lies in the orthogonality, or "Neyman orthogonality", of the moment condition. Orthogonality implies that small estimation errors in the nuisance components $m(X)$ and $e(X)$ have minimal influence on the consistency and asymptotic normality of the treatment effect estimator $\hat{\tau}(x)$. Formally, orthogonality requires that the moment function has zero derivative with respect to the nuisance parameters at their true values. This condition ensures robustness to regularization bias introduced during the estimation of high-dimensional or complex nuisance functions.

In the estimation equation A.8, this orthogonality is achieved by subtracting the estimated nuisance components from the observed outcome and treatment assignment. Specifically, by replacing y_i with $y_i - \hat{m}(X_i)$ and A_i with $A_i - \hat{e}(X_i)$, the estimator isolates variation in y that is unrelated to confounding due to X_i. This transformation yields a debiased estimating equation whose expectation conditional on X is centered at zero at the true value of $\tau(X)$, i.e.,

$$\mathbb{E}\left[\left(y_i - m(X_i) - \tau(X_i)(A_i - e(X_i)) \right) (A_i - e(X_i)) \right] = 0,$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

for all X_i. This moment condition satisfies the orthogonality requirement, meaning that the estimator $\hat{\tau}(x)$ is *locally insensitive* to first-stage estimation error. As a result, double machine learning methods like the causal forest are valid and efficient even when the nuisance functions are estimated flexibly via machine learning.

A1.1.1. First-Stage Estimation of Nuisance Functions

In the first stage of estimation, the goal is to flexibly estimate the nuisance functions $m(X) = \mathbb{E}[y \mid X]$ and $e(X) = \mathbb{E}[A \mid X]$, representing the expected outcome and treatment propensity given covariates. These functions are crucial for constructing an orthogonalized estimating equation in the second stage. To accommodate the high-dimensional and potentially nonlinear relationships in the data, we use machine learning methods, specifically, random forests to estimate $m(X)$ and $e(X)$. A random forest regressor is used for $m(X)$, while a random forest classifier is used for $e(X)$.

These models rely on a set of values known as *hyperparameters*, which govern the structure and flexibility of the model. Unlike parameters learned from the data during training (like coefficients in a regression), hyperparameters are preset before training begins and control aspects such as model complexity, overfitting, and generalization.

In this implementation, the two key hyperparameters tuned for each nuisance function are:

- **Maximum tree depth** (`max_depth`): This limits how deep each decision tree in the forest can grow. Larger values allow the model to fit more complex relationships but increase the risk of overfitting. We consider $\texttt{max_depth} \in \{10, 20, 30\}$.
- **Minimum sample split** (`min_samples_split`): This specifies the minimum number of samples required to split an internal node. Larger values lead to simpler trees and smoother predictions. We consider

$$\texttt{min_samples_split} \in \{1000, 2000, 5000\}$$

.

These hyperparameters directly influence the bias-variance trade-off. A small `min_samples_split` and large `max_depth` may lead to overfitting the noise in the training data, while the reverse can underfit and fail to capture

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

important signal. To select appropriate hyperparameter combinations, We perform cross-validation, using out-of-sample prediction error to balance this trade-off.

We use 100 trees for estimating the nuisance functions. This number balances estimation stability and computational efficiency. Random subsampling of data and averaging across trees reduce variance and provide robust estimates of $\hat{m}(X)$ and $\hat{e}(X)$. Importantly, the feature set used for $e(X)$ includes only variables available at the time of impression (e.g., website and timestamp), consistent with the platform's real-time decision-making process.

A1.1.2. Second-Stage: Estimating Treatment Effects via Causal Forests

In the second stage, the Conditional Average Treatment Effect (CATE), denoted by $\tau(x)$, is estimated using the causal forest algorithm. The causal forest builds on the random forest framework but modifies its splitting rule and objective function to focus on heterogeneity in treatment effects rather than minimizing prediction error.

Each causal tree in the forest is trained on a random subsample of the data. Within each subsample, the algorithm recursively splits the observations based on values of the covariates X to maximize heterogeneity in treatment effects between the resulting subgroups. The splitting criterion for each node is designed to maximize the sum of squared differences in estimated treatment effects across the child nodes:

$$\max_{S_1,S_2} \ \hat{\tau}_1^2 + \hat{\tau}_2^2,$$

where $\hat{\tau}_1$ and $\hat{\tau}_2$ denote the estimated treatment effects in the two child nodes S_1 and S_2. This criterion ensures that the splits create partitions with maximally distinct treatment effects, leading to more informative subgroup-specific estimates.

After growing many such trees, the causal forest aggregates the results across trees to compute a smooth, nonparametric estimate of the local treatment effect for each observation. Specifically, the importance weight $\alpha_i(x)$ measures the similarity between an observation i and a target point x, and is proportional to the number of trees in which both i and x fall into the same leaf. These weights act as a kernel, giving greater influence to nearby or similar observations when estimating $\hat{\tau}(x)$.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

As with the first-stage models, the causal forest requires a set of hyperparameters that govern the structure and complexity of the forest. Hyperparameters control how deeply trees grow, how finely the sample space is partitioned, and the trade-off between bias and variance. The specific values tuned for the causal forest are:

$$\texttt{max_depth} \in \{10, 20, 30\}, \quad \texttt{min_samples_split} \in \{1000, 2000, 5000\}.$$

A deeper tree (larger `max_depth`) allows for capturing more complex heterogeneity in treatment effects, but may overfit if too deep relative to sample size. Conversely, a larger `min_samples_split` results in smoother, more stable estimates at the cost of reduced flexibility.

These hyperparameters are chosen via cross-validation to minimize out-of-sample error in treatment effect estimation. In this analysis, we use 300 trees to estimate the treatment effects. Increasing the number of trees reduces the variance of the estimator by averaging over multiple randomized partitions, improving the stability of the resulting $\hat{\tau}(x)$ estimates.

After training, the importance weights $\alpha_i(x)$ are calculated as the proportion of trees in which both i and x fall within the same terminal leaf:

$$\alpha_i(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{L_t(X_i) = L_t(x)\},$$

where $T = 300$ is the total number of trees and $L_t(\cdot)$ denotes the leaf assignment in tree t. These weights form the basis for the locally weighted regression that estimates $\hat{\tau}(x)$, with higher weights assigned to observations that are most similar to x in the covariate space.

A1.2. Final Prediction and Counterfactual Estimation

With nuisance functions and weights estimated, the Conditional Average Treatment Effect (CATE) can be computed using a locally weighted regression (Equation A.8). Additionally, the model allows estimation of counterfactual click probabilities using:

$$\hat{\mathbb{E}}y(X_i|A = a) = \hat{m}(X_i) + \hat{\tau}(X_i) \cdot (A_a - \hat{e}(X_i)).$$

where A_a is defined by equation A.2. This representation comes from the identity:

$$\mathbb{E}[y^a|X] = m(X) + \tau(X) \cdot (A - e(X)),$$

which decomposes the observed outcome into a baseline prediction and a treatment effect deviation from the mean assignment probability.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

A2. More Details on Inputs to the Click Probability Model

A2.1. Historical User Characteristics

A2.1.1. Browsing History

User browsing history can reveal some information about user demographics and preferences for different ads. For example, in the United States, internet users who visit news websites are more likely to be Millennials, male, high-income, and have a college degree[1].

According to Table 2 in the paper, the websites are classified into 13 groups based on their subject. We use 13 indicator variables for each user's previous visit to those groups to account for their browsing histories.

In Figure A2.1, users are classified into eight distinct groups based on their browsing behavior on the three most popular website subjects (politics, health, and finance), and the average click rates of these user groups on four ads are depicted. Significant heterogeneity is observed across different user types. For example, Ad 1 achieves the highest click rate among users who have only visited health websites. At the same time, Ad 3 performs best among those who have not visited any of the categorized websites, and Ad 4 has the highest click rate among those who just visited financial websites.

A2.1.2. Previous Exposures to the Ad

Several studies have highlighted the decreasing returns to repeated ads. To account for this effect, the number of user exposures to each ad since the beginning of the week is included in the set of variables used for estimating click rates. Figure A2.2 depicts the average click rates for the most frequently shown ad in the data versus the number of times it has been repeated to the user since the beginning of the week (including the current repeat). This figure suggests that the probability of clicking on this ad drops when it gets more repeated.

A2.1.3. Previous Clicks on the Ad

Having previously clicked on an ad suggests that the user is interested. So they are more likely to click on it again. This effect is demonstrated in Figure A2.3 for the most frequently shown ad in the data. This figure indicates that users who have previously clicked on an ad once and twice are

[1] https://www.statista.com/study/119126/target-audience-news-website-readers-in-the-united-states/

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

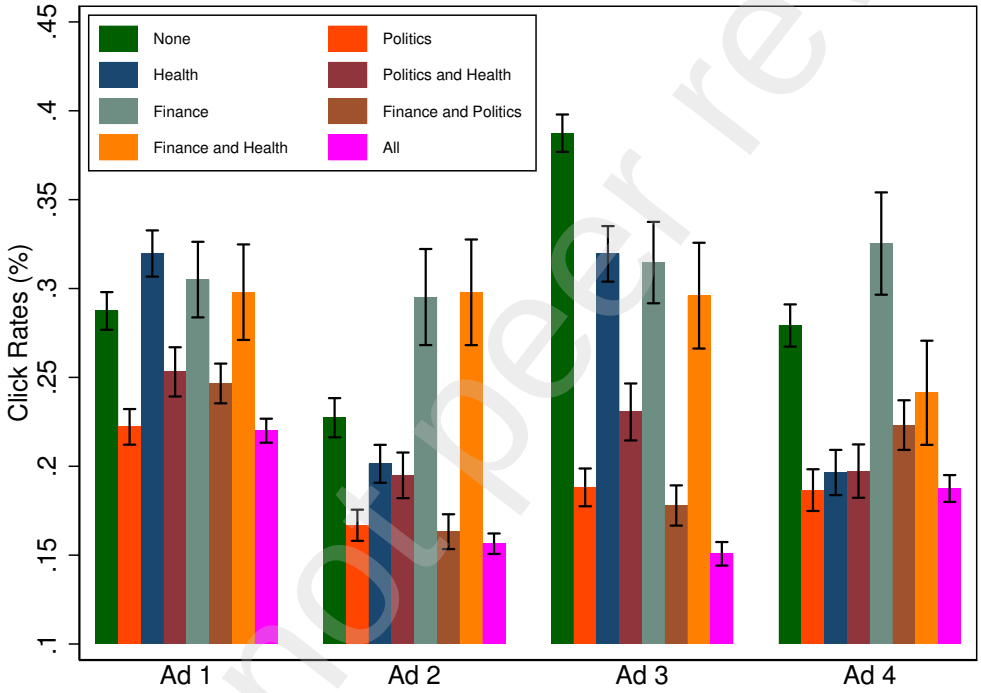


Figure A2.1: Average User Click Rates Based on What Type of Websites They Have Visited

Notes: The calculations are based on one week of data from September 3-9, 2022, for top 4 ads in the data. Average click rates and their 95% confidence interval are depicted for different user groups.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

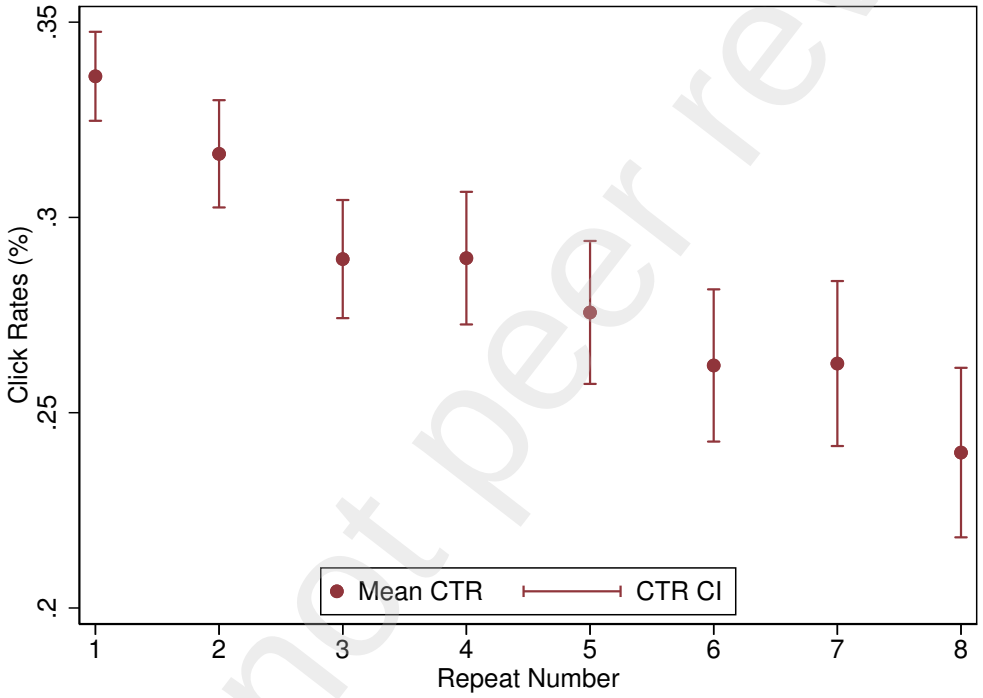


Figure A2.2: Average User Click Rates Versus The Number of Repeats

Notes: Average click rates and their 95% confidence intervals are depicted for different impression repeats. Repeat numbers are the number of times the user has visited the same ad since the beginning of the week (including the current impression). The calculations are based on one week of data from September 3-9, 2022, for the most frequently shown ad in the data.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

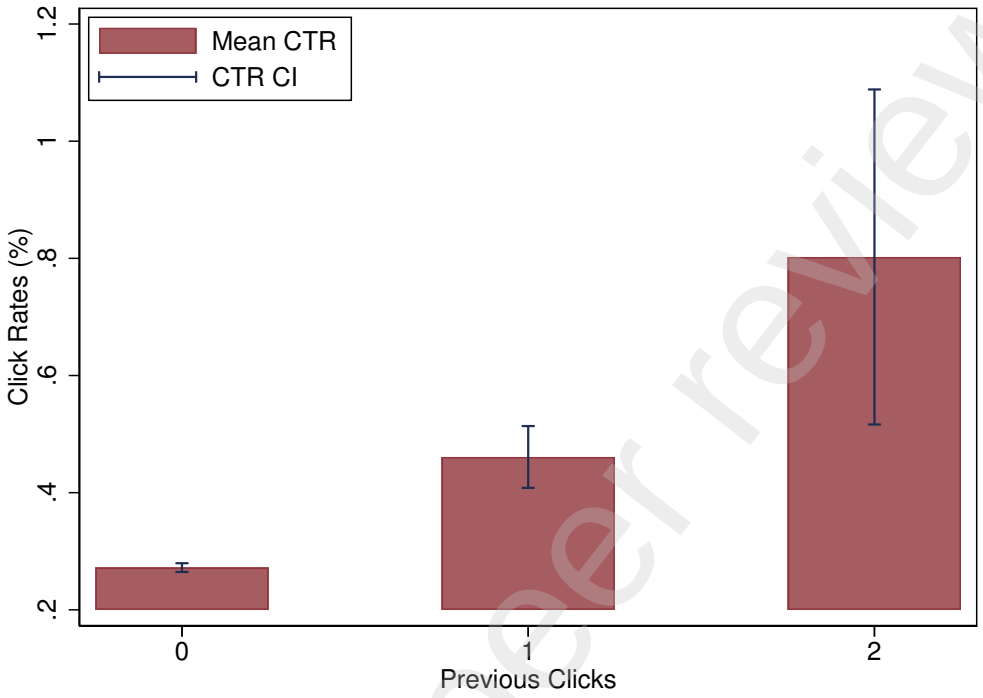


Figure A2.3: Average User Click Rates Versus The Number of Previous Clicks

Notes: Average click rates and their 95% confidence interval are depicted for different numbers of previous clicks since the beginning of the week. The calculations are based on one week of data from September 3-9, 2022, for the most frequently shown ad in the data.

twice and three times more likely to click on that again later relative to those who have never clicked, respectively.

A2.1.4. User Activity Level

Generally, users who are more active online are less sensitive to ads and less likely to click. This effect can be seen in Figure A2.4, which shows the average click rate of users on the most frequently shown ad based on the number of pages they have visited. Users who have visited ten pages or less are three times more likely to click on this ad relative to those who have visited more than 100 pages.

A2.1.5. Current Website Subject

Different types of websites attract users with varying interests and ad preferences for ads. This heterogeneity can be shown in Figure A2.8 which

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

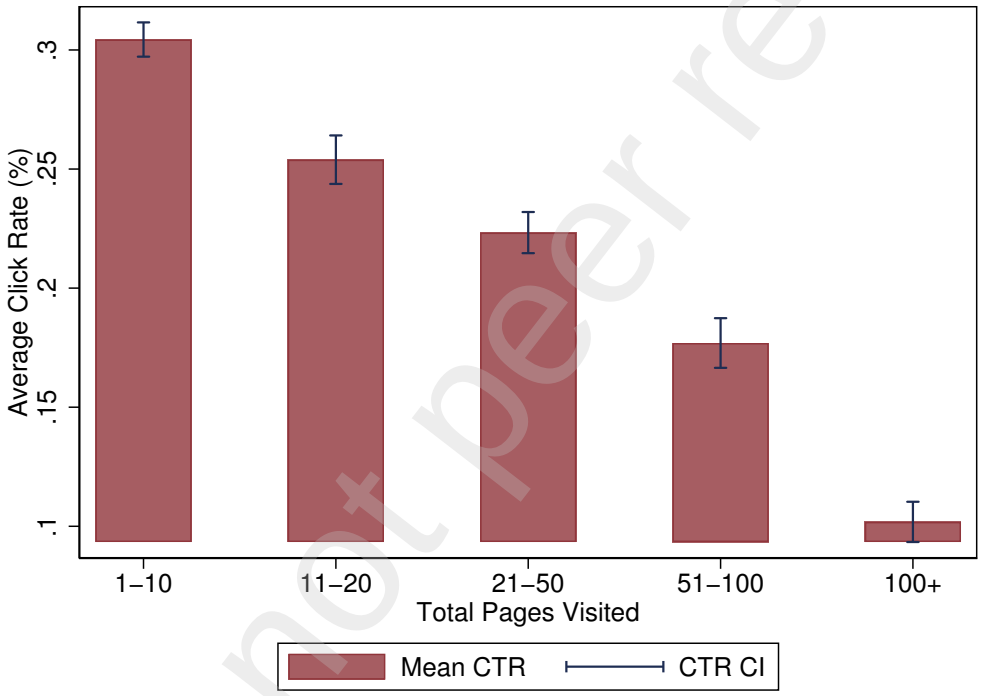


Figure A2.4: Average User Click Rate Across Users with Different Activity Levels

Notes: This figure depicts the average click rates and their 95% confidence intervals for the top ad across users with different activity level, measured by the total number of pages they previously visited.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

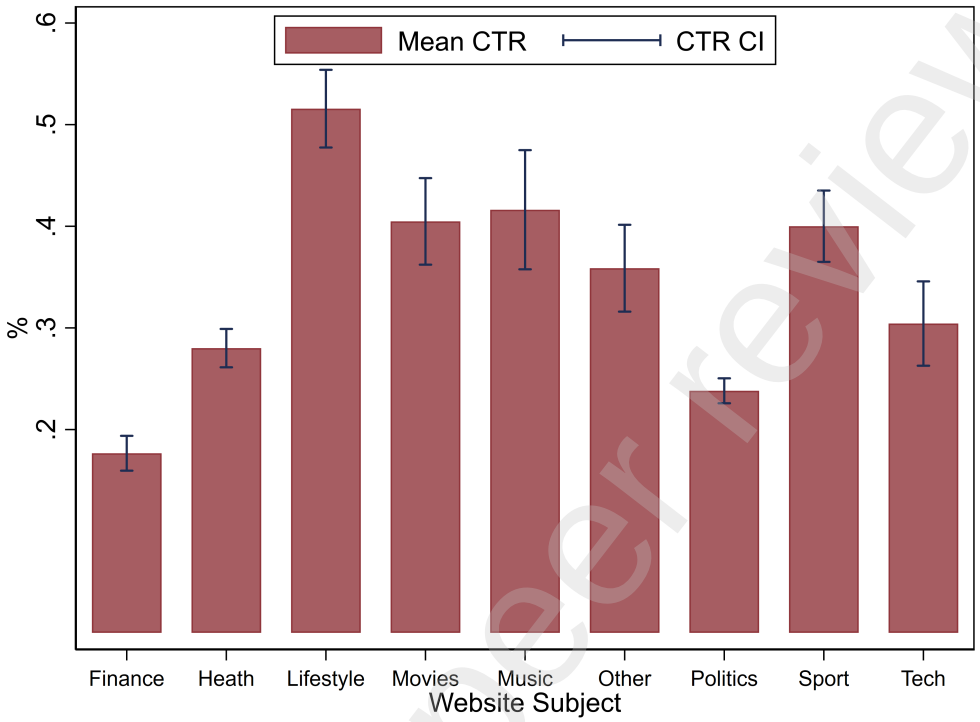


Figure A2.5: Average User Click Rates Across Websites with Different Subjects
Notes: This figure depicts the average click rates for the top ad on different websites based on their subjects and their 95% confidence intervals.

shows the average click rate for the most frequently shown ad in the data across different website subject groups. The average click-through rate for this ad is nearly three times higher on Lifestyle websites compared to Finance websites.

To account for this heterogeneity, we include 13 indicator variables representing the subject matter of the current website.

A2.1.6. Within-subject rank of the current website

Ads have different average click rates on different websites, even within the same website subject group. This is shown in Figure A2.6, where the average click rate of the most popular ad in the data is plotted on different websites within the same subject (political). To control for that, we include an additional variable for the rank of the current website within the websites with the same subject. Since the method used for estimating the conditional

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

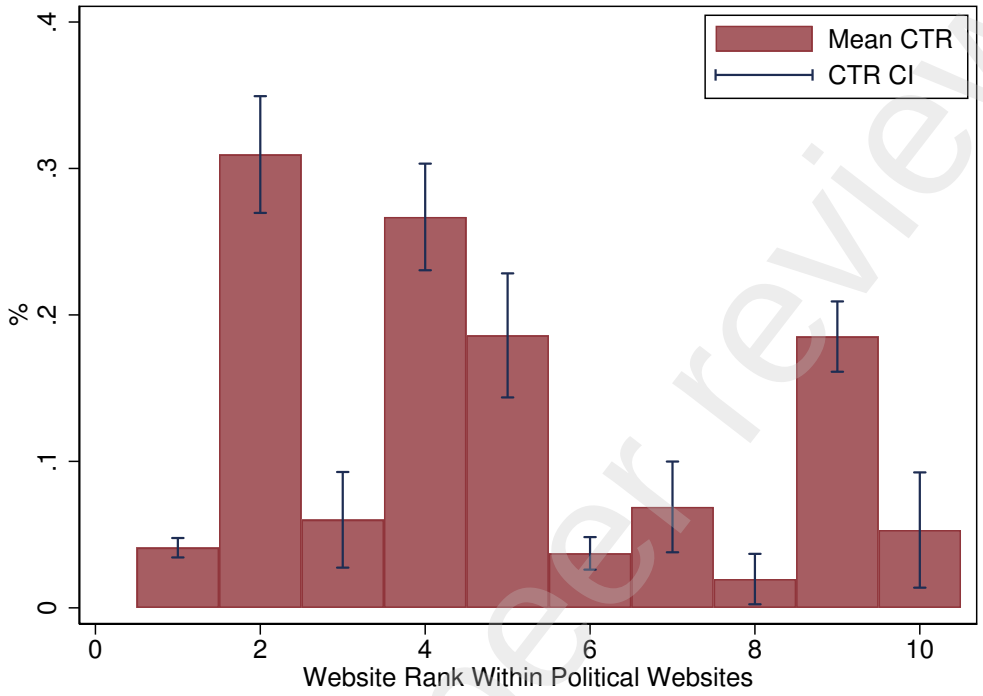


Figure A2.6: Average User Click Rates Across Different Political Websites

Notes: This figure depicts the average click rates for the top ads on the most visited political websites and their 95% confidence intervals.

click rates is non-linear, this variable and the current website fixed effect can jointly control for the current website fixed effect.

A2.1.7. Number of ad slots on the page

The higher the number of ad slots on a page, the less user attention they will receive and the lower the click rate. This effect is shown in regressions provided in Table A2.1. In these regressions, each observation is the average click rate of each of the top 100 ads on a website. The dependent variable is the logarithm of the average click rate of the ad on the website, and the independent variable of interest is the average number of ads on the website. Four specifications with various fixed effects (FEs) for the advertiser, website subject, and website are run. The results in Table A2.1 show that the number of ads on a page negatively impacts the average click rate, with a 7-13% reduction per additional ad.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4
Num. Ads on Page	-0.112***	-0.129***	-0.0683***	-0.0781***
	(0.00812)	(0.00672)	(0.00783)	(0.00717)
Constant	-4.403***	-4.295***	-4.689***	-4.632***
	(0.0555)	(0.0479)	(0.0542)	(0.0492)
Observations	4,489	4,489	4,489	4,447
R-squared	0.053	0.268	0.329	0.572
Advertiser FE	NO	YES	YES	YES
Publisher Subject FE	NO	NO	YES	YES
Publisher FE	NO	NO	NO	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A2.1: the Effect of Number of Ad Slots per Page on Average Click Rates
Notes: The dependent variable is the log of the average click rate of each of the top 100 ads on each website. The data for the last day of the data is only used. The independent variable is the average number of ads displayed per page for each website. In model 4, publisher rank within subject is used as a fixed effect.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

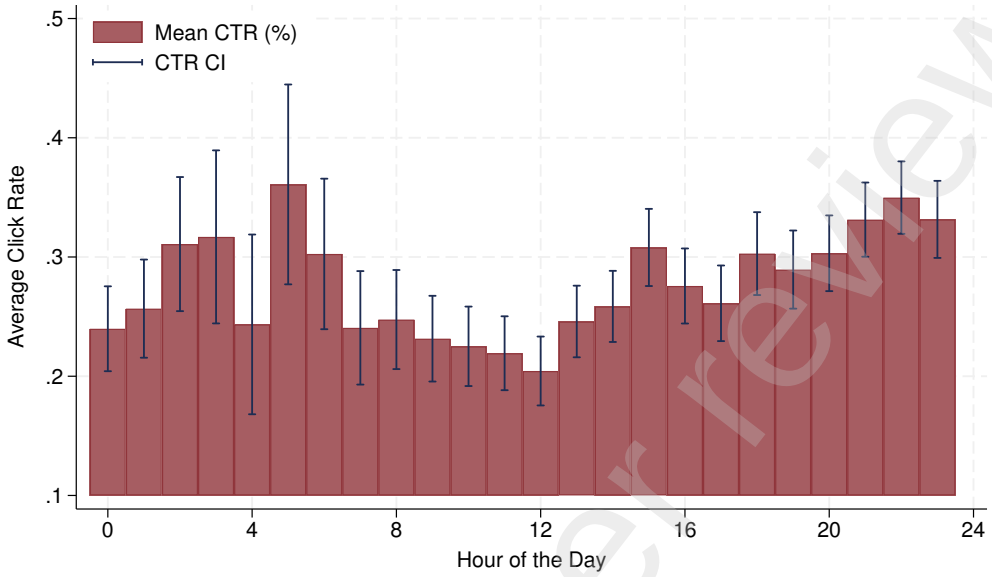


Figure A2.7: Average User Click Rate in Different Hours of the Day

Notes: This figure depicts the average click rates and their 95% confidence intervals for the top ad across different times of the day. The last two days of data are used for the estimates.

A2.1.8. Time of Impression

As discussed in the paper, time affects the propensity scores. In addition, it could affect the click rates since different types of users may browse websites at different times. Figure A2.7 depicts the average click rate for the most frequently shown ads in the data and its 95% confidence interval for different hours of the day. As can be seen, users are more likely to click on this ad in the evenings than in the mornings.

To control for this effect, we add a fixed effect for days and a continuous variable for the hour of the day in the model.

A2.1.9. User Device Type

Generally, ads have significantly lower click rates among desktop devices. The average click rates for the top 10 ads in the data among mobile and desktop devices are depicted in Figure A2.8.

A2.1.10. Zero Propensity Scores

For each ad among the top 100 ads in the data (which are the ads included in the main analysis), we compute how many websites among the top 200

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

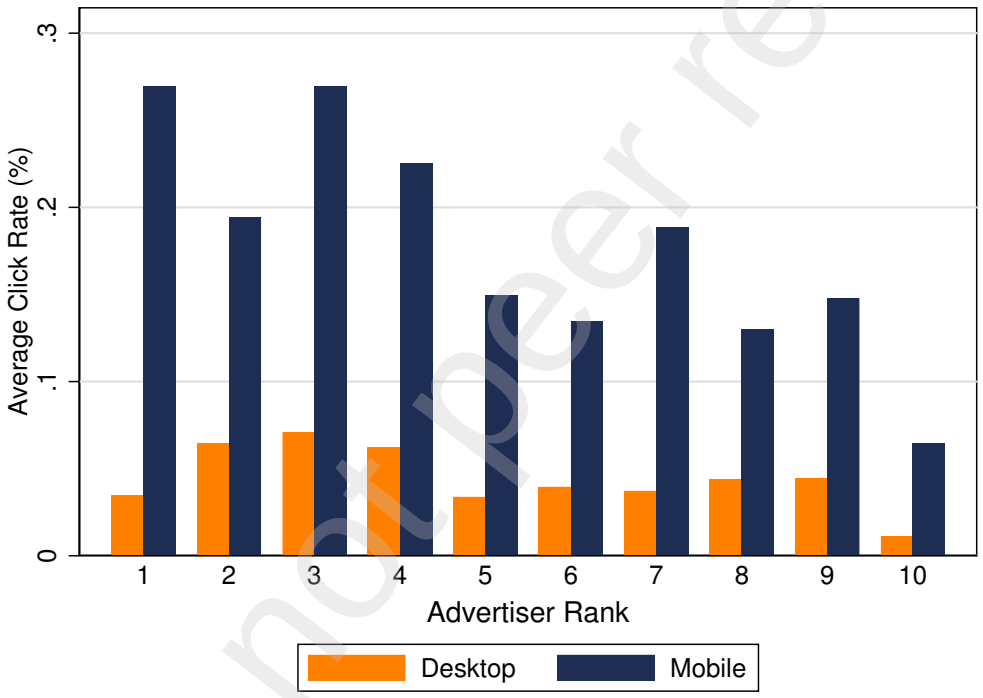


Figure A2.8: Average User Click Rates Among Desktop and Mobile Devices

Notes: This figure depicts the average click rates for the top ten ads across desktop and mobile devices.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

never display the ad. This provides a measure of zero propensity scores across publishers. Figure A2.9 presents a histogram of these counts across ads. As shown in the figure, approximately 60% of the ads are displayed on all top 200 websites. Additionally, about 15% of the ads are missing from only one website, and no ad is missing from more than eight websites. These results suggest that most ads are widely distributed across publishers in the data.

While a zero empirical propensity score may seem concerning from an identification standpoint, it does not necessarily imply that the true underlying propensity score is zero. Rather, it could result from a low but nonzero propensity combined with a limited number of observations. In other words, even if an ad is eligible to be shown on a particular website, it may not appear in the sample due to randomness or small sample size.

Nevertheless, the key insight from this figure is that the data provide substantial support across ad–website combinations. There is no evidence that advertisers selectively restrict their ads to a small set of websites, which would violate the overlap assumption required for causal inference. This widespread distribution alleviates concerns about support violations and strengthens the credibility of the identification strategy used in the estimation of treatment effects.

A3. Recovering Advertiser Valuations

Advertisers only derive utility when users click on their ads, not merely when their ads are displayed. This assumption is particularly relevant for small or unknown advertisers, who rely on click-throughs to generate brand awareness or sales. If users do not engage with the ad, they are unlikely to recall the product or the advertiser. In practice, advertisers often care about ultimate conversions, such as purchases or sign-ups. However, because we do not observe post-click outcomes, we use click behavior as a proxy for advertiser utility, which is a common approach in the online advertising literature.

To recover advertiser valuations, we invert the observed bids under the assumption that advertisers are bidding optimally to maximize their expected utility. An advertiser a receives utility only if their ad is both shown and clicked. If the advertiser bids b_a, the expected utility can be written as:

$$U_a = (v_a - b_a) \underbrace{\mathbb{P}\left(A_i = a\right)}_{\text{prob. of ad } a \text{ being shown}} \underbrace{\mathbb{P}\left(y_{ia} = 1 | A_i = a\right)}_{\text{prob. of ad } a \text{ receiving a click}} \tag{C.1}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

The probability of ad a being shown is determined by the platform's allocation rule. Letting y_a denote the probability that ad a receives a click, we can rewrite the utility as:

$$U_a(b_a; y_a) = (v_a - b_a) \times e(b_a y_a) \times y_a \tag{C.2}$$

where $e(.)$ is the propensity score function: the probability of ad a being shown conditional on the bid and predicted click rate. The first-order condition with respect to b_a is:

$$(v_a - b_a)\, e'(b_a y_a) y_a = e(b_a y_a) \tag{C.3}$$

where $e'(.)$ is the derivative of the propensity function with respect to b_a.

Assuming a constant allocation rule across websites and times, and one bid and click rate per advertiser, we write $e(b_a y_a)$ as:

$$e(b_a y_a) = \frac{b_a y_a}{\sum_{a' \in A_{wt}} b_{a'} y_{a'}} = \frac{b_a y_a}{b_a y_a + \sum_{a' \in A_{wt}, a' \neq a} b_{a'} y_{a'}} \tag{C.4}$$

Therefore:

$$e'(b_a y_a) = \frac{y_a \times \left(\sum_{a' \in A_{wt}, a' \neq a} b_{a'} y_{a'} \right)}{\left(b_a y_a + \sum_{a' \in A_{wt}, a' \neq a} b_{a'} y_{a'} \right)^2} = \frac{e(b_a y_a)\,(1 - e(b_a y_a))}{b_a} \tag{C.5}$$

by plugging in C.5 into C.3, the advertiser valuations can be recovered:

$$v_a = b_a + \frac{e(b_a y_a)}{e'(b_a y_a)} = b_a + \frac{b_a}{1 - e(b_a y_a)} \tag{C.6}$$

Figure A3.1 depicts the histogram of the estimated advertiser valuations.

A4. Shapley Values

Shapley values, originally proposed in cooperative game theory, provide a principled way to attribute a model's prediction to its input features. The idea is to treat each feature as a "player" in a game whose "payout" is the model's prediction. The Shapley value of a feature is then the average marginal contribution of that feature to the prediction across all possible subsets of features.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Formally, let F denote the set of all features and $S \subseteq F \setminus \{i\}$ be a subset not containing feature i. The Shapley value $\phi_i(X)$ for feature i at instance X is defined as:

$$\phi_i(X) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[\tau_{S \cup \{i\}}(X) - \tau_S(X)\right]$$

Here, $\tau_S(X)$ is the treatment effect estimated using the features in set S, and the weighting term ensures fair attribution over all possible subsets.

A5. Feature Importances

While Shapley values provide a local, observation-specific measure of variable contribution to treatment effects, the feature importance scores offer a global view of how features drive heterogeneity across all ads and all users. To assess the global importance of each feature across all treatment effect estimates, we compute tree-based feature importance scores. This approach quantifies how much each variable contributes to the heterogeneity of treatment effects by evaluating its role in partitioning the data during tree construction.

Let $FI_{as}(x_k)$ denote the importance of feature x_k at split level s in a tree estimating the treatment effect for ad a. It is computed as:

$$FI_{as}(x_k) = \frac{w_{\text{parent}} \cdot w_{\text{left}} \cdot w_{\text{right}}}{w_{\text{parent}}^2} \cdot (\tau_{\text{left}} - \tau_{\text{right}})^2$$

Here, w_{parent} is the number of observations in the parent node, w_{left} and w_{right} are the counts in the left and right nodes after the split, and τ_{left}, τ_{right} are the treatment effects in each child node. This score captures how much heterogeneity the split on feature x_k introduces.

The cumulative feature importance for feature x_k across all splits with depth $d_s \leq 4$ in the tree for ad a is:

$$FI_a(x_k) = \sum_{s:\text{depth}_s \leq 4} \frac{1}{(1 + \text{depth}_s)^2} \cdot FI_{as}(x_k)$$

To obtain a single importance score across all ads, we average the feature importance over the set of ads:

$$FI(x_k) = \frac{1}{|A|} \sum_{a \in A} FI_a(x_k)$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

A6. More Details on the Duopoly Counterfactuals

The duopoly counterfactuals aim to evaluate how different data splits across competing platforms affect ad allocation decisions, user experience, and market outcomes. To that end, we consider three intuitive scenarios that vary in how they assign access to impression and click data across two platforms. These scenarios are:

1. **Split by Websites:** In this counterfactual, each platform is responsible for showing ads on half of the total websites in the network. The split is implemented to ensure that the number of observations available to each platform is roughly balanced. Specifically, we first sort the websites in descending order based on their total number of visits. Then, to construct a balanced allocation, we assign the websites ranked 1st, 4th, 5th, 8th, and so on to Platform 1, and the websites ranked 2nd, 3rd, 6th, 7th, and so on to Platform 2. This alternating scheme ensures both platforms are exposed to a comparable volume of user activity, while still being restricted to different parts of the web ecosystem.
2. **Split by Website Subjects:** In this setup, we partition the set of websites based on their subject categories, with each platform being assigned responsibility for specific topics. The goal is again to keep the number of user interactions roughly balanced across the two platforms. To achieve that, we rank the website subjects by total visit volume and assign subjects with ranks 1, 4, 5, 8, and so on to one platform, and those ranked 2, 3, 6, 7, and so on to the other. The result is a content-based segmentation of the digital ad space. The final assignment of subjects is as follows:

   - **Platform 1:** Art, Education, Entertainment, Health, Lifestyle, Politics, Sports
   - **Platform 2:** Childcare, Finance, Movies, Music, Other, Technology

   This partition is designed to emulate environments where platforms are specialized in particular content domains, such as a health-focused media network versus a finance-oriented platform, and can only learn user preferences through their assigned content categories.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

3. **Split by Users:** In the third counterfactual, we assign users randomly to one of the two platforms. Each platform can show ads to, and collect data from, only their own group of users. In contrast to the other two splits, both platforms can access the full set of websites and subjects but are limited to observing behavior from only a subset of the population. This scenario mimics competitive environments where platforms are differentiated not by content or publishers, but by their user bases.

Table A6.1 reports summary statistics for the data platforms used in estimating click-through rates (CTR) in each counterfactual scenario. For the monopoly case, the platform has access to the full dataset, which includes approximately 27.2 million ad impressions, 1,850 unique websites, and over 800,000 users. In the duopoly scenarios, each platform is limited to a subset of the data based on the specific split logic (by website, website subject, or user), and therefore has access to fewer observations. For each split, the table shows the number of impressions, websites, and users that each platform sees, along with the corresponding average CTRs.

The table highlights that the overall coverage and user access vary significantly depending on the type of split. For example, in the website subject split, Platform 1 ends up with more observations but lower CTR (0.148), while Platform 2 receives fewer observations but significantly higher CTR (0.211), suggesting variation in user engagement across website categories. In contrast, the user and website splits result in fairly balanced platforms with similar CTRs. This comparison provides early evidence of how information fragmentation across platforms could affect ad allocation quality and user targeting precision in the counterfactual simulations.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

Preprint not peer reviewed

Counterfactual	Platform No.	Observations	Websites	Users	Average CTR (%)
Monopoly	-	27,156,537	1,850	802,120	0.170
Random Split by Website	1	13,385,499	920	534,485	0.170
	2	13,771,038	930	594,585	0.169
Random Split by Website Subject	1	17,840,260	636	538,867	0.148
	2	9,316,277	1,214	524,027	0.211
Random Split by User	1	13,578,696	1,736	401,215	0.170
	2	13,577,841	1,716	400,905	0.169

Table A6.1: Characteristics of Data Platforms Used for Click Rate Estimation in Different Counterfactuals

Notes: Websites and users measure the number of unique websites and users in each platform's dataset. Observations correspond to the last two days of data used for CTR estimation. CTR stands for click-through rate.

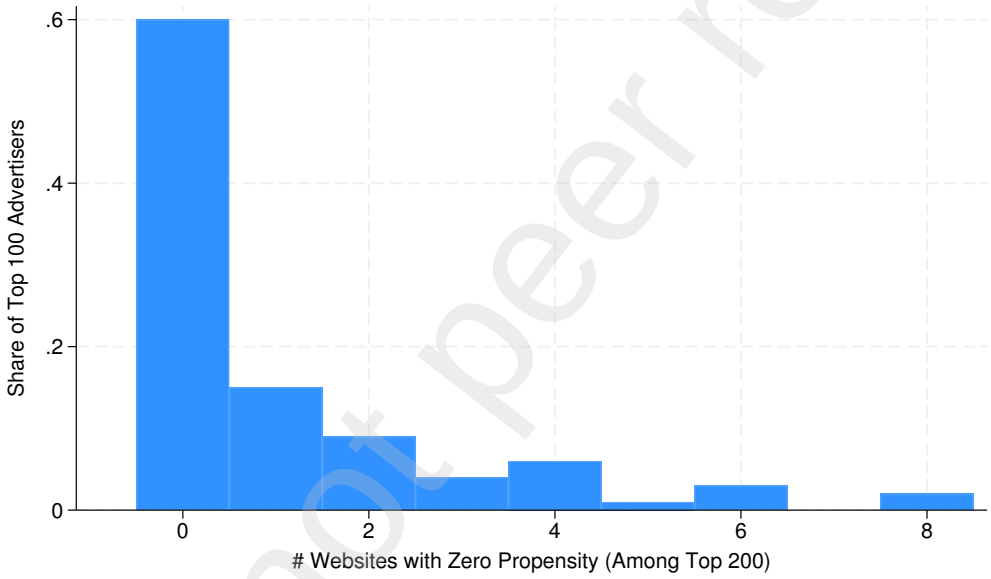


Figure A2.9: Histogram of the Number of Websites Top 100 Ads Are not Shown on

Notes: This figure depicts the number of websites each of the top 100 ads is not shown on. The calculations are based on one week of data from September 3-9, 2022.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076

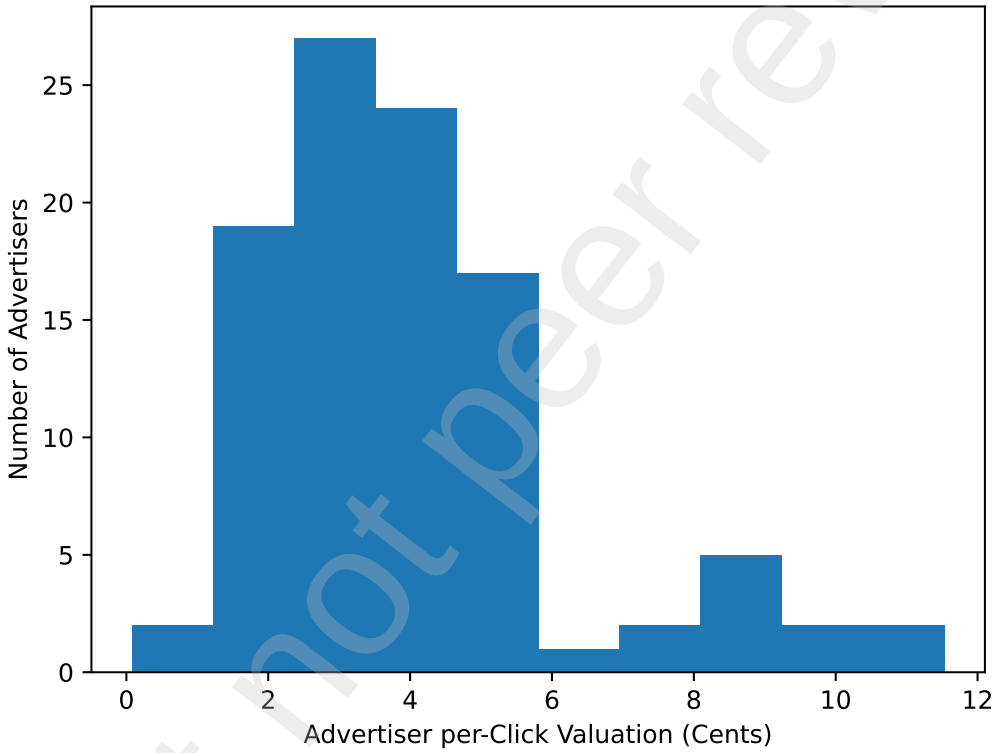


Figure A3.1: Histogram of Advertiser Valuations

Notes: Valuations are shown for the top 100 advertisers in the data. Valuations are converted to cents using the average exchange rate in September 2022.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=6504076