The Effect of Ad Blocking on User Engagement with the Web

Ben Miroglio
Mozilla Corporation
Mountain View, CA, USA
bmiroglio@mozilla.com

David Zeber
Mozilla Corporation
Mountain View, CA, USA
dzeber@mozilla.com

Jofish Kaye
Mozilla Corporation
Mountain View, CA, USA
jkaye@mozilla.com

Rebecca Weiss
Mozilla Corporation
Mountain View, CA, USA
rweiss@mozilla.com

ABSTRACT

Web users are increasingly turning to ad blockers to avoid ads, which are often perceived as annoying or an invasion of privacy. While there has been significant research into the factors driving ad blocker adoption and the detrimental effect to ad publishers on the Web, the resulting effects of ad blocker usage on Web users' browsing experience is not well understood. To approach this problem, we conduct a retrospective natural field experiment using Firefox browser usage data, with the goal of estimating the effect of adblocking on user engagement with the Web. We focus on new users who installed an ad blocker after a baseline observation period, to avoid comparing different populations. Their subsequent browser activity is compared against that of a control group, whose members do not use ad blockers, over a corresponding observation period, controlling for prior baseline usage. In order to estimate causal effects, we employ propensity score matching on a number of other features recorded during the baseline period. In the group that installed an ad blocker, we find significant increases in both active time spent in the browser (+28% over control) and the number of pages viewed (+15% over control), while seeing no change in the number of searches. Additionally, by reapplying the same methodology to other popular Firefox browser extensions, we show that these effects are specific to ad blockers. We conclude that ad blocking has a positive impact on user engagement with the Web, suggesting that any costs of using ad blockers to users' browsing experience are largely drowned out by the utility that they offer.

CCS CONCEPTS

• **Information systems** → **Online advertising**; **Browsers**; • **General and reference** → *Empirical studies*;

KEYWORDS

Ad Blocking; Propensity Scoring; Matching; Natural Field Experiment; User Studies

1 INTRODUCTION

For an average user, a typical day on the Web involves exposure to ads. Indeed, advertising has become the primary revenue model for many popular websites, most notably search engines, media outlets and streaming services. There has been substantial research into user perception of online ads and the steps taken to avoid them. Ads are often seen as annoying, or lead to a negative Web browsing experience [7, 10, 15], and the prevalence of behavioral or retargeted ads raises concerns about privacy [20, 27, 49]. This in turn leads to reduced effectiveness from the point of view of ad publishers [21, 25].

As a result, Web users are increasingly turning to ad blocking to mitigate the negative effects of online ads. A recent study estimates that over 600M devices worldwide were using ad blocking by the end of 2016, of which over half were mobile; this represents a 30% increase since 2015 [42]. Furthermore, ad blockers top the list of most popular Firefox extensions, with at least 18M installs [34].

1.1 The effect of ad blocking on the Web ecosystem

Ad blocking is, understandably, very unpopular among ad publishers, with the IAB recently labelling the practice a "potential existential threat" [8]. This is quantified by a recent industry study which found that the use of ad blocking leads to significant worsening in the traffic rank of sites that depend on ads [48], a compound effect of the revenue loss. Ad publishers' attempts to combat ad blockers by limiting site access to ad blocker users has lead to an "arms race" [39], although the net effect is that users generally abandon such sites [42].

On the other hand, there are a number of ways that employing ad blockers improves users' experience on the Web, beyond avoiding the annoyance of ads. For one thing, the amount of data loaded when visiting pages with ads is significantly reduced [18, 45], leading to savings in both load times and data costs on mobile. In fact, among a selection of popular news sites, over half the data loaded (in aggregate) was found to be ad-related [2]. In addition to boosting site performance, blocking ads reduces exposure to privacy and

security threats associated with ads such as behavioral tracking and malvertising [14, 19, 28]. As well as being measurable quantitatively, these benefits are understood by users, as demonstrated by a number of user research studies on this topic [4, 30, 41, 42].

1.2 What is the effect of ad blocking on Web engagement?

Although much of the recent research has focused on the specific mechanisms by which ad blockers impact user experience or the economic impact to ad publishers, the overall effect of users' choice to use an ad blocker on their general Web usage has not been well studied. This is the contribution we make in this work, by addressing the question: how different is user engagement with the Web for users who installed an ad blocker in their browser from those who did not? If ad blocking is ultimately detrimental to the Web ecosystem as a whole, we expect to see a drop in Web usage among ad blocker users. On the other hand, if alleviating the negative experience caused by ads actually outweighs the potential breakage, we hypothesize that Web usage would be increased.

We tackle this problem by studying Firefox usage data. The Firefox browser reports daily aggregate statistics on Web usage, such as total active time in the browser and number of pages loaded, as well as currently installed browser extensions ("add-ons"). Drawing on a historical dataset of Firefox usage, we frame the question as a natural experiment [13].

Restricting to new users who installed Firefox during a specific time period, we select a test group of users who installed an ad blocker, and a control group that did not. The selection is performed using a matching technique applied to baseline user features, which allows us to infer a degree of causality from the observed effects. We then compare test group Web usage after installing the ad blocker to control group usage over a comparable period, controlling for baseline usage levels. This is sometimes referred to as analysis of covariance, and is closely related to difference-in-difference analysis.

There are many benefits afforded by this approach. By working with usage data as reported by the browser from a large set of users, we obtain a large-scale, "objective" view that doesn't depend on the specific ways in which blocking ads modifies users' experience on the Web or on subjective user responses. Additionally, all of the required data is already collected by the Firefox browser as a part of normal operation. This means that there are no additional costs in terms of participant recruitment or data collection. Furthermore, the design can be revised and the analysis rerun as needed with little additional overhead. In fact, we make use of this feature to run placebo studies for the usage effects of other Firefox add-ons, to further bolster the causality claim.

Moreover, designing the study as a natural experiment allows us to estimate the effect of actually making the choice to use an ad blocker. In contrast, a randomized experiment where users are randomly assigned to use an ad blocker, while offering the strongest guarantee of causality, can only measure the consequences of blocking ads on the Web; users' discovery of ad blocking technology and their choice to use it plays no role. For this reason, natural experiment designs are commonly used when studying the effects

of users' choices or behaviors, such as in the context of social media and online communities [3, 5, 11, 16], or online advertising [9, 17].

2 EXPERIMENTAL DESIGN

The goal of our study is to quantify the causal effect of ad blocking usage on user engagement with the Web via the Firefox browser. The preferred methodology for causal inference, the so-called "gold standard", is a randomized experiment, where users are randomly assigned to test and contol groups. In our context, that would mean that a random subset of users get an ad blocker installed in their browser.

However, there are two main reasons why this approach is infeasible in our setting. For one thing, forcibly installing an ad blocker, or any other extension which would substantially change browsing experience, in users' browsers raises ethical concerns for Mozilla, as it may not respect users' preferences and may adversely affect their browsing experience.

The other issue is that a randomized design lacks *ecological validity*, i.e. it does not reflect the way ad blocking is generally used. In most cases, a user comes to learn about ad blocking technology, often via word of mouth [4, 42], and then makes a decision to install an ad blocker to address a bad Web experience. Additionally, there are likely systemic differences between users who learn about ad blockers (or browser extensions in general) and further choose to act on that knowledge, and those who never do. Also, because of the prevalence of ads, the Web experience after installing the ad blocker becomes noticeably different from the user's point of view, which they may associate with their choice to use an ad blocker. This in turn may affect their subsequent usage. A randomized experiment cannot take these effects into account. Instead, the randomization "removes" any implicit differences in the population groups by ensuring that they are balanced.

Phrasing our research question in more general terms, we are interested in the value to users of having ad blockers available as part of the browser extension ecosystem, measured in terms of browser usage. To this end, we opt for causal inference via a natural experiemnt (i.e. an observational study [46]). It is important to note that the causality we are seeking to assert is not that *ad blocking will affect any Firefox user's Web usage in the observed way* but rather that *the effects we observe among the test group are due to their installation of an ad blocker and not other implicit factors*. Any fundamental differences between ad blocker users and the rest of the population are not addressed in these findings.

2.1 Data

In this study, we analyze usage data reported by users of the desktop Firefox Web browser. During the course of typical browser usage, Firefox sends usage data to Mozilla using a data collection system referred to as Telemetry [36]. The exclusive data source for this study is a Telemetry data record known as the *"main" ping* [35], which is sent at least once a day for each active user, and contains a range of activity and environmental measurements such as:

- overall and active browser session time
- numbers of searches (initiated from the UI search bars) and pages loaded
- numbers of tabs and windows opened

- whether Firefox is the default browser
- the current default search engine
- date of Firefox installation
- app information, e.g. version and distribution channel
- system information, e.g. OS and version, memory size
- the list of current browser add-ons.

Each record is annotated with a unique user ID and submission date, which allows for user-level longitudinal analysis.

This study is purely retrospective, in the sense that all the data we use had already been collected prior to its inception as a part of normal operation of the native client. We do not perform any kind of recruitment or specialized data collection.

2.2 Study periods

Among the users represented in the Telemetry dataset, we restrict to those located in the United States who installed the standard English-language (en-US locale) build of Firefox during the month of February 2017. This choice of population segment helps to reduce extraneous variability due to regional differences in Web browsing behaviour and seasonality in browser usage, which is generally consistent from February up until the summer months.

For each of these users, we then segment their first 70 calendar days since installation into three consecutive periods:

- a *baseline period* covering the first 28 days
- a *treatment period* consisting of the next 14 days
- an *observation period* lasting the remaining 28 days.

We restrict to users who installed no add-ons during their baseline period, and require that they submitted at least 5 Telemetry records during each of the baseline and observation periods, to ensure that we have sufficient data. We then select a *test* group of users who installed any add-on at some point during the treatment period, and a *control* group who installed no add-ons during any of their three periods. We will further segment test users according to which add-on it was that they installed, with a primary focus on ad blockers, as described in Section 3.2.

We adopt a non-equivalent groups design, in which we compare browser usage during the observation period between the groups, controlling for features recorded during the baseline period. We allow 28 days for both baselining and observation so that we can assess users' established browser patterns both before and after installing the add-on.

Note that, because each user's timescale begins when they install Firefox, their periods may span different calendar dates. Nonetheless, any two users' time periods are offset by no more than 27 days, since they are constrained to have installed Firefox during February 1-28. These relationships are illustrated in Figure 1.

Selected in this way, our dataset contains 16,414 test users and 342,528 control users—our test group being much smaller due to the additional restrictions we imposed. We now proceed to select an appropriate subset of the control group via matching.

2.3 Matching procedure

Since group membership is based on observed behaviour and not assigned by randomization, it may be the case that users in the test group are just more engaged to begin with, rather than their experience being affected by the use of an ad blocker. To remove this

Figure 1: Illustration of overlapping user study periods
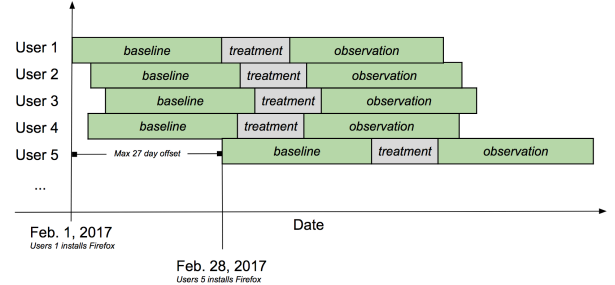
contingency and to reinforce the causality in our results, we seek to ensure that the groups are *a priori* balanced on any observable features, i.e. that there are no systematic differences between them. This is achieved using matching. Each test group member is paired with a similar member of the control group, where similarity is determined based on features observed during the baseline period. This ensures that the groups start off "more or less the same", e.g. that one group is not biased towards more engaged users.

2.3.1 Propensity scoring. Similarity between users is assessed using propensity scores, which represent a user's likelihood to install an add-on. Such a score can be computed using a logistic regression model which predicts the probability a user will install an add-on, given what we observe about them in the baseline period.

More formally, for each user i, let $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$ denote a vector of k features observed during the baseline period, and let T_i indicate group membership:

$$T_i = \begin{cases} 1 & \text{if user } i \text{ is in the test group} \\ 0 & \text{otherwise} \end{cases}$$

The propensity score for user i is defined as

$$\pi_i := \pi(X_i) = Pr(T_i = 1 \,|\, X_i).$$

Scores are computed by fitting a logistic regression model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = X_i\beta + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

to yield estimates $\hat{\pi}_i = g(X_i\hat{\beta})$, where g is the inverse of the logit function used in the LHS of (1). The list of features used in propensity scoring is provided in Table 1.

2.3.2 Correction for class imbalance. From a modeling perspective, we have a class imbalance problem, which can bias our model by heavily weighting majority (control) class features, dwarfing any separable features in the minority (test) class. To mitigate this, we undersample from the control group, drawing a random sample of control users equal in size to the test group to create a balanced dataset. However, in the process, we risk losing valuable information due to the exclusion of the majority of the control group.

We correct for this by means of a bootstrap aggregation approach. Let n_T be the number of test users. We draw m samples of size n_T from the control group. Each sample is combined with the full test group to create a dataset of size $2n_T$, and the logistic model (1) is

fit to this reduced dataset to obtain propensity score estimates $\hat{\pi}_{ij}$, $i = 1, \ldots, n, j = 1, \ldots, m$. The final propensity score we use in matching is obtained by averaging across the replicate model fits:

$$\hat{\pi}_i := \frac{1}{m} \sum_{j=1}^{m} \hat{\pi}_{ij}. \tag{2}$$

Figure 2 shows that the test group, on average, was assigned higher propensity scores than the control group, meaning that the baseline features X_i do in fact contain information about a user's eventual decision to install an add-on. This distributional difference between the propensity scores of our two groups justifies the need for the matching procedure, since the groups are not directly comparable otherwise.

Figure 2: Distribution of propensity scores by group ($m = 1000$)
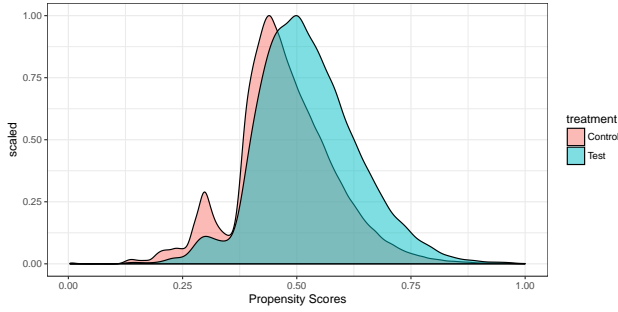
2.3.3 Matching. Each test user is matched to the control user whose propensity score is closest to its own, provided the difference is less than a set threshold δ, otherwise that test user is dropped from the dataset. In other words, given a test user u_t, we find the set of candidate matches

$$C(u_t) = \left\{u_c \in \text{control} : |\hat{\pi}_{u_c} - \hat{\pi}_{u_t}| \leq \delta\right\}.$$

If $|C(u_t)| = 0$, u_t is not matched, and is excluded from further consideration. Otherwise, we select the control user u_c satisfying

$$\text{argmin}_{u_c \in C(u_t)} |\hat{\pi}_{u_c} - \hat{\pi}_{u_t}|,$$

and retain the pair of users.

Note that reducing δ improves the balance of the dataset at the cost of reducing its size. We ran matching for various values of δ, and found that setting $\delta = 0.0001$ retains over 99% of the test group. This is the value we use in the following.

Also, it is important to note that a single control user can be matched to multiple test users using the procedure described above. We account for this in subsequent modeling by weighting each record by the inverse of its frequency (i.e. if a control user occurs 3 times in the matched dataset, we assign it a weight of $\frac{1}{3}$). In the final matched dataset, duplication is rare: approximately 2% of control users occur twice, and less that 1% occur three times. The final dataset contains 32,825 records (16,414 and 15,724 **distinct** test and control users, respectively).

Table 1: Features used for propensity scoring (X)

Is default	Boolean indicating if Firefox is the user's default browser
Operating system	Windows, MacOS, Linux
Browser version	Firefox version string
Memory MB	Memory size on a user's machine
Sync configured	Boolean indicating if a user set up a sync account
Active hours	Total time the user spent interacting with the browser
Session length	Total browser session time
Number of subsessions	Total Telemetry reports
Total tab open events	Number of tabs opened
Total URI count	Total number of pages loaded
Unique URI count	Number of unique websites (TLD+1) visited
Address bar searches	Number of searches initiated from the address bar
Search bar searches	Number of searches initiated from the dedicated search bar
Total Yahoo searches	Number of searches made using the Yahoo search engine
Total Google searches	Number of searches made using the Google search engine
Total searches	Total number of searches overall initiated from one of the Firefox search bars
Number of bookmarks	Total bookmarks added
Length of history	Total number of pages visited
Quantum ready	Boolean indicating if a client qualifies for the Firefox Quantum Project
E10s enabled	Whether multiprocess is enabled
E10s cohort	Branch assignment for the multiprocess experiment

2.4 Match evaluation

After applying the matching procedure, it is important to validate that the groups are indeed indistinguishable in terms of the k covariates observed in the baseline period, as this will influence the selection of the final model. We treat categorical and continuous features separately in this evaluation.

2.4.1 Categorical features. If matching was successful, categorical variables should have approximately the same relative frequencies between the test and control groups. In other words, the distribution of the categorical variable should be independent of the group label. For each categorical feature listed in Table 1, we test independence between the groups using a Chi-square Test of Independence. We apply the test to the dataset both before and after matching, for comparison purposes.

Table 2 lists the P-values resulting from each test run. Recall that a small P-value indicates there is dependence between the groups (relative frequencies are different). At the 5% significance level ($\alpha = 0.05$), we fail to find dependence on the group label for any categorical feature observed in the matched dataset. However, prior to matching, each P-value is very close to 0, implying that the baseline period frequency distributions do differ between the groups.

Table 2: Chi-square test P-values before/after matching

Feature	Before	After
Is default	0	0.3242
Operating system	0	0.3124
Browser version	0	0.9908
Sync configured	0	0.1237
Quantum ready	0	0.5410
E10s enabled	0	0.6606
E10s cohort	0	0.6138

2.4.2 Continuous features. Similarly, we check for distributional differences between the test and control groups for the continuous features. As these distributions tend to be highly skewed and suffer from discreteness, we apply two nonparametric tests, which in conjunction give us good sense of the "balance" in the dataset. We compare ECDFs using a bootstrapped Kolmogorov-Smirnov (KS) test, for which the P-value is computed empirically over 1000 bootstrapped samples, to provide coverage for discreteness [1]. Additionally, we use the Quadratic-Chi Histogram Distance (QCHD) [43] to compare the raw histograms, again bootstrapping the P-value (as a permutation test).

Table 3 shows the resulting P-values for both tests applied to the data before and after matching. Again, a small P-value indicates lack of independence between the groups. As evidenced by the low P-values, we have not balanced the continuous features as well as the categorical features. We will handle this imbalance by controlling for a user's baseline activity in the final models.

3 MODELING AND RESULTS

We now estimate the effect of installing an ad blocker on observation period Web engagement.

3.1 Engagement response measures

User engagement is an abstract concept described in terms of users' cognitive involvement with a technology, which generally relates to the quality of their experience [24]. While the direct study of user engagement typically employs qualitative research methodology [6, 40], it is common to quantify engagement with a product or website using proxy measurements that capture the amount in which users interact with the technology, such as usage or dwell time, page loads, clicks, etc [26, 29, 44]. In this study, we consider three such measures of Web engagement.

Active hours roughly captures the amount of mouse and keyboard interaction with the browser. It is based on a Telemetry

Table 3: Bootstrap KS test and QCDH permutation test P-values before/after matching

Feature	KS		QCDH	
	Before	After	Before	After
Memory MB	0	0	0	0
Active hours	0	0	0	0
Session length	0.001	0.016	0.228	0
Number of subsessions	0	0	0	0
Total tab open events	0	0	0	0.561
Total URI count	0	0.007	0	0.037
Unique URI count	0	0	0	0.019
Address bar searches	0	0.006	1	1
Search bar searches	0	0	1	1
Total Yahoo searches	0	0.014	1	1
Total Google searches	0	0	1	1
Total searches	0	0	1	1
Number of bookmarks	0	0.008	0	0.01
Length of history	0	0.002	0	0.215

measurement called *active ticks*. The browser session is split into 5-second intervals, and each interval which saw key presses or mouse movements is counted as an active tick. [1] We consider this measurement an indicator of Web usage in cases where a given website or Web application demands interactivity from the user.

Total URIs is a counter that reflects the total amount of requests that the browser makes in a given session. We consider this variable to serve as an indicator of Web navigation, regardless of the interactivity of the websites a particular client visits.

Finally, we also consider **Search Volume** made using the search bars embedded in the Firefox UI (also referred to as *search access points*). We included this variable as an enagement measure since we assert that search is a fundamental part of modern Web navigation for most users today.

The combination of these variables allow for us to analyze engagement across a broad array of usage patterns on the Web. For example, the active hours response might be very useful for users who tend to visit highly interactive websites, but would underestimate Web usage for users who primarily consume streaming content.

3.2 Model specification

We estimate differences in Web enagement by fitting linear regression models of the form

$$\log(Y_{obs}) = \beta_0 + \beta_1 \log(Y_{base}) + \beta_2 T + X\gamma + \epsilon, \quad (3)$$

where:

- Y_{obs} is one of Active Hours, Total URI Count, or Search Volume as measured during the observation period,
- Y_{base} is the baseline value for the same measure,

---

[1]For example, a user that moves the mouse in the browser for 60 consecutive seconds will register 12 active ticks.

- T is a treatment indicator (taking the value 1 if the user belongs to the treatment group and 0 otherwise), and
- $X = (X_1, X_2, \ldots, X_v)$ is a vector of continuous features recorded during the baseline period.

Each of the engagement measures is log-transformed in order to correct for high skewness, a common characteristic of such Web-related measurements. The continuous features X are included in the model as controlling factors, since these were not fully balanced during the matching process as noted in Section 2.4.2.

It is important to note that, although our matching procedure incorporates the baseline measures Y_{base}, none of these is adequately balanced after the process. Nonetheless, since we are modelling relative differences per user along these measures, an imbalance for Y_{base} between treatment and control is acceptable since it is explicitly controlled for. Additionally, continuous features X are included in the model (3) to ensure that any imbalances unresolved by matching do not skew the estimate for β_2. For example, when $Y = $ *Active Hours*, we include all continuous features to better isolate the true effect that T has on Y_{obs}. In fact, the rounded estimates in Table 5 remain unchanged if X is excluded from the model, however we include it to bring attention to its role in the model design.

Although the primary focus of this study is the installation of ad blockers, we can redefine the test group to contain users who installed any specific set of Firefox add-ons, maintaining the control group of users who did not install any add-on. We can then rerun the entire procedure outlined in previous sections and refit model (3) with the updated treatment indicator T. Following this approach, we consider four distinct test groups, determined by which add-on was installed during the treatment period (defined in Section 2.2). The add-ons for each group are laid out in Table 4.

Table 4: Add-ons for test groups

| Test group	Add-ons installed
adblocker	either AdBlock Plus (ABP) [32] or uBlock Origin [37], the two most popular ad-blocking Firefox add-ons
any_addon	any Firefox add-on [33]
vdh	Video DownloadHelper (VDH), a popular add-on for downloading videos from streaming sites [38]
msa	McAfee WebAdvisor (MSA), a security focused add-on that blocks malicious sites and downloads [31]

The *adblocker* group is the test of primary interest. The additional test groups, still compared against a control group with no add-ons, serve as placebo tests to verify that the effect we observe for ad blockers is indeed distinguishable from the act of installing add-ons in general, a popular add-on, or a security-focused add-on. Given the breadth of the Firefox add-on ecosystem, it would be desirable to perform a number of other such placebo tests for different types of add-ons. However, because of the restictions imposed by the study design, a user belongs to a given treatment group only if we observe an group-specific add-on installation in the treatment

period *after* a 28-day baseline period. Such specificity limits the range of viable placebo candidate add-ons present in our study sample. Among the available placebo candidates, we selected VDH as the most popular general-interest add-on, and MSA as the top security-focused add-on (noting that "Privacy & Security" is the most popular "specialized" category of add-ons).

3.3 Results

For each response measure for Web engagement, and each test group listed in Table 4, we fit model (3). The main parameter of interest is β_2, which represents the change in the response measure due to installing the relevant add-ons, given similar baseline usage levels. In particular, inverting the log transformation, $\exp(\beta_2) - 1$ gives the relative change in the average value of the engagement measure during the observation period, relative to the control group, controlling for baseline period activity. The fitted values of β_2 for each of the models are presented graphically in Figure 3, together with 95% confidence intervals.

Figure 3: Estimates and 95% confidence intervals for β_2, the change in log-transformed engagement due to installing add-ons
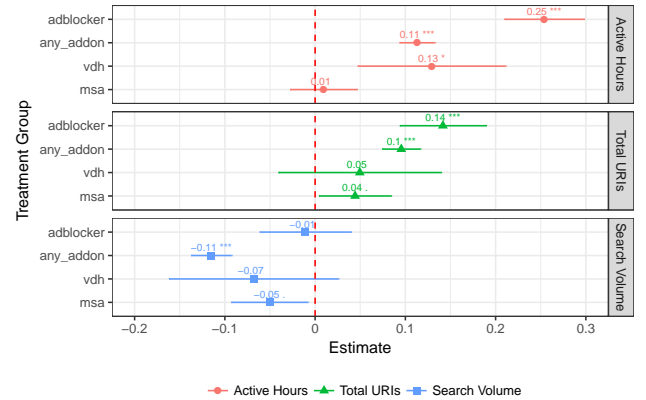
Similarly, the estimated relative changes in the engagement measures are summarized in Table 5.

Table 5: Estimated relative changes in engagement due to installing add-ons compared to control group ($\exp(\beta_2) - 1$)

Measure	adblocker	any_addon	vdh	msa
Active Hours	28%	11%	14%	0%
Total URIs	15%	10%	0%	5%
Search Volume	0%	−12%	0%	−5%

Taking into account the matching and adjustments made to ensure comparability between the groups, we find that installing an ad blocker does in fact have a strong effect on average Web engagement. Users who installed an ad blocker were active in the browser for around 28% more time on average, and loaded 15% more pages (URIs), controlling for baseline activity.

These results are reinforced by comparing them against the other test groups. Indeed, the notable differences in the estimates reflect the diversity present in the add-ons ecosystem, and highlight the sensitivity of our engagement measures to the specific interactions of users with each individual add-on, beyond any common baseline effect of just "having an add-on". In particular, users who installed *any* add-on saw an average increase of around 10% in both active hours and page loads. This may be explained in part by the fact that the user took steps to personalize their browser, and consequently remains more engaged. However, the effect on page loads for ad blocker users is stronger, and considerably so for active hours, taking into account the margins of error indicated by the confidence intervals. This suggests that there is something inherent to the experience of using an ad blocker that proves beneficial to the user, increasing their engagement with the Web via the browser.

This conclusion is further borne out by the effects for the VDH and MSA tests, which, while both involving popular add-ons that provide a specific beneficial functionality, fail to affect engagement to the degree observed for ad blockers. One plausible explanation is that the use of an ad blocker provides a noticeably enhanced experience for users in that group as they navigate their favorite Web pages, leading to increased engagement in terms of both active time and page views. On the other hand, Video DownloadHelper provides a specialized utility on specific Web sites (downloading videos from streaming sites), and the presence of McAfee WebAdvisor only surfaces in rare occasions (visits to malicious sites). The results for these test groups show little to no increase in Web engagement, and certainly not more than the baseline case of installing any add-on. An interesting direction for future research would be to investigate whether the principle holds in general: how well is an individual add-on's effect on Web engagement via the browser explained by the specific way in which it interacts with the user experience?

There are a few additional observations worth highlighting. None of the test groups for specific add-ons appears to have a significant effect on search volumes, whereas the overall effect of installing a general add-on is to reduce search volumes by around 10%. This is not surprising considering that the specific add-ons we tested do not generally interact with search functionality, whereas there are a number of general add-ons related to search that likely draw searches away from the search bars in the browser UI. Additionally, Figure 3 shows notable differences in the margins of error between the different test groups. This is also not surprising, and is largely explained by the differences in sample sizes between the groups. The distribution of add-on popularity, as measured by their number of users, is highly skewed [34]. As a result, among the top few most popular add-ons, popularity drops by an order magnitude between consecutive ranks, and there is a long tail of numerous add-ons with very few users at the other end of the list.

4 DISCUSSION

We now address some of the limitations inherent in this experimental approach, and describe some interesting directions for future investigation.

4.1 Methodological limitations

While the results are substantially significant from a statistical perspective, we note that this approach is not without limitations. Perhaps the most obvious issue is that our study only incorporates usage data from Firefox users. Thus, while the results generalize to the population of Firefox users, they may not apply to users of other browsers, such as Chrome. However, Firefox users would have to be substantially different from other browser users in their usage of the Web in order to challenge the generalizability of our findings. While this may be the case, it seems unlikely in light of the contemporary state of the Web. The majority of Web traffic goes to a limited set of popular websites, and the nature of Web standards entails that there are limited differences between browsers in terms of core Web browsing operations. Therefore, if there are in fact differences between Firefox and Chrome users, say, they are more likely related to latent user-level preferences and beliefs rather than typical Web usage. In other words, the choice of Firefox over Chrome is more likely explained by brand preference than the likelihood to use popular websites.

Another limitation is that in order to allow a sufficiently long baseline period, we only consider users that installed an ad blocker after their first 28 days as a Firefox user. This selection process introduces certain biases into the applicable population. In our internal anlyses, we observe that Firefox users who install add-ons generally do so within their first week of using the browser. Requiring that users install an ad-blocker at least a month after their first usage instance means restricting to a specific subpopulation, and may introduce bias in terms of other features (e.g. lower than average general activity levels). On the other hand, we have also observed that Firefox users who eventually stop using the browser tend to do so during their first few weeks. This suggests that our study results are biased towards users more likely to use Firefox consistently.

We also must consider that, although this is a retrospective study, it is not long-term, in the sense that we are not observing users longer than about 2 months. We are not able to take into account long-term activity, and the effects we observe may be related to the novelty of the experience. To elaborate, it may be entirely possible that over a longer period of observation, user engagement with the Web eventually dwindles back down to pre-treatment levels or worse. However, this limitation can be empirically addressed through future research where we adjust the observation period lengths.

Finally, matching according to propensity scores allows for the approximation of random assignment for observational studies. There is an assumption, however, that features not included in the propensity scoring model do not possess explanatory power for response variables of interest. In other words, it is possible that there are potentially unobserved variables that we either failed to include or that the browser is not in a position to measure (e.g. latent variables endogenous to users) that could factor into the likelihood of installing extensions. That being said, this limitation is standard for natural field experiments, which are ultimately not a true replacement for randomized controlled trials.

4.2 Future research directions

There are a number of ways we may consider expanding on this current work and addressing some of the limitations described above. For one thing, the design of the study depends on parametrized assumptions, such as the lengths of the periods described in Section 2.2. To strengthen the results, it would be valuable to assess their robustness to changes in these parameters.

One question to investigate is how the results change over time. In the current study, we observe users over 28 days after the add-on installation and aggregate their usage measurements over this period. Alternatively, we could extend the observation period to span several months, and adopt a more sophisticated modeling approach that would take into account changes over time as well as user churn.

Another issue raised above is that the fixed baseline period length requires restricting the population to users who install an add-on during a specific time window in their lifetime. We could instead shrink the baseline period and extend the treatment period, in order to capture a broader segment of the user population. However, a shorter baseline period may reduce the effectiveness of matching and diminish the causal claims. This tradeoff will need to be considered carefully in order to arrive at an optimal result.

Additionally, we may derive more detailed insights by further developing the models describe in Section 3.2. In the current work we opted for relatively simple models for ease of interpretability and presentation. However, it would be of interest to allow for more sophisticated relationships between the features, such as an interaction between the treatment effect and baseline period engagement (i.e. does the difference during the observation period depend on how engageed the users started off as during the baseline period?). Another improvement would be to combine the separate models into a single, larger model, which would let us pool variance between different test groups and potentially analyze multivariate relationships between the engagement measures. However, as this type of data is typically not "well-behaved" in a traditional statistical sense (it suffers from skewness and discreteness), this may require additional transformations or adjustments to yield a meaningful result.

Finally, as discussed in Section 2, a randomized experiment will always allow for stronger causal claims, although it may be difficult to implement, and may in fact address a somewhat different research question. That said, it would be very interesting to run such an experiment to see how the results compare to our current findings. As noted in Section 2, there may be ethical implications to requiring study participants to use an ad blocker, although this could be resolved by making the study opt-in, for example. Of course, this again incurs a tradeoff in that the stronger degree of causality comes at the expense of population bias. Alternatively, we could consider targeting users in the current study with surveys in order to understand their actual motivations and experiences around ad blocker usage. Tied back to their reported Telemetry data, this would provide an additional verification on the sensitivity of our causal results.

5 CONCLUSION

While the impact of ad blocking on the ecosystem of the Web has been substantially investigated and documented, there has been little research into the impact of ad blocking on engagement with the Web. Opponents of ad blocking have asserted that ad blocking results in substantial breakage of modern websites and Web applications, which results in poor user experience and decreased Web engagement. This outcome, they suggest, could threaten the future growth of the Web. On the other hand, proponents of ad blocking have argued that ads are such a detriment to the browsing experience that users are willing to make the tradeoff by removing ads from their browsing experience entirely.

In this paper, we present a natural field experiment that addresses this tension directly by estimating the causal effect of installing ad blocking extensions on various measures of Web engagement. We find that installing ad blocking extensions substantially increases both active time spent in the browser and the number of pages viewed. This empirical evidence supports the position of ad blocking supporters and refutes the claim that ad blocking will diminish user engagement with the Web.

REFERENCES

[1] Alberto Abadie. 2002. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *J. Amer. Statist. Assoc.* 97, 457 (2002), 284–292.

[2] Gregor Aisch, Wilson Andrews, and Josh Keller. 2015. The Cost of Mobile Ads on 50 News Websites. (October 2015). Retrieved October 26, 2017 from https://www.nytimes.com/interactive/2015/10/01/business/cost-of-mobile-ads.html

[3] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 537–546. https://doi.org/10.1145/3018661.3018672

[4] Mimi An. 2016. Why People Block Ads (And What It Means for Marketers and Advertisers). (July 2016). Retrieved October 27, 2017 from https://research.hubspot.com/reports/why-people-block-ads-and-what-it-means-for-marketers-and-advertisers

[5] Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature Communications* 8 (2017), 14753.

[6] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. 2011. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*. 9–12.

[7] Giorgio Brajnik and Silvia Gabrielli. 2010. A review of online advertising effects on the user experience. *International Journal of Human-Computer Interaction* 26, 10 (2010), 971–997.

[8] Interactive Advertising Bureau. 2017. IAB Believes Ad Blocking Is Wrong. (2017). Retrieved October 25, 2017 from https://www.iab.com/iab-believes-ad-blocking-is-wrong/

[9] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 7–16. https://doi.org/10.1145/1835804.1835809

[10] Chang-Hoan Cho and Hongsik John Cheon. 2004. Why do people avoid advertising on the internet? *Journal of Advertising* 33, 4 (2004), 89–97.

[11] Tiago Cunha, Ingmar Weber, and Gisele Pappa. 2017. A Warm Welcome Matters!: The Link Between Social Feedback and Weight Loss in /R/Loseit. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1063–1072. https://doi.org/10.1145/3041021.3055131

[12] Stylianos Despotakis, R Ravi, and Kannan Srinivasan. 2017. The Beneficial Effects of Ad Blockers. (2017). http://stedes.com/pdfs/DRS%20-%20The%20Beneficial%20Effects%20of%20Ad%20Blockers%20-%202017.pdf

[13] Thad Dunning. 2012. *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084444

[14] Catherine Dwyer and Ameet Kanguri. 2017. Malvertising - A Rising Threat To The Online Ecosystem. *Journal of Information Systems Applied Research* 10, 3 (2017), 29–37.

[15] Steven M Edwards, Hairong Li, and Joo-Hyun Lee. 2002. Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of Advertising* 31, 3 (2002), 83–95.

[16] Seyed Amin Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the Effect of Exercising on Users' Online Behavior. In *The Workshops of the Eleventh International AAAI Conference on Web and Social Media.* AAAI Technical Report WS-17-16.

[17] Gian M Fulgoni and Marie Pauline Mörn. 2009. Whither the click? How online advertising works. *Journal of Advertising Research* 49, 2 (2009), 134–142.

[18] Kiran Garimella, Orestis Kostakis, and Michael Mathioudakis. 2017. Ad-blocking: A Study on Performance, Privacy and Counter-measures. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17).* ACM, New York, NY, USA, 259–262.

[19] Arthur Gervais, Alexandros Filios, Vincent Lenders, and Srdjan Capkun. 2017. Quantifying web adblocker privacy. In *European Symposium on Research in Computer Security.* Springer, 21–42.

[20] Avi Goldfarb and Catherine Tucker. 2011. Online display advertising: Targeting and obtrusiveness. *Marketing Science* 30, 3 (2011), 389–404.

[21] Daniel G Goldstein, Siddharth Suri, R Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. 2014. The economic and cognitive costs of annoying display advertisements. *Journal of Marketing Research* 51, 6 (2014), 742–752.

[22] James J. Higgins. 2003. *Introduction to Modern Nonparametric Statistics* (1 ed.). Cengage Learning.

[23] Frank J. Massey Jr. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78.

[24] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. 2014. *Measuring User Engagement.* Morgan & Claypool Publishers.

[25] Anja Lambrecht and Catherine Tucker. 2013. When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research* 50, 5 (2013), 561–576.

[26] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of User Engagement. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12).* Springer-Verlag, Berlin, Heidelberg, 164–175. https://doi.org/10.1007/978-3-642-31454-4_14

[27] Wen Li and Ziying Huang. 2016. The Research of Influence Factors of Online Behavioral Advertising Avoidance. *American Journal of Industrial and Business Management* 6, 09 (2016), 947.

[28] Zhou Li, Kehuan Zhang, Yinglian Xie, Fang Yu, and XiaoFeng Wang. 2012. Knowing Your Enemy: Understanding and Detecting Malicious Web Advertising. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12).* ACM, New York, NY, USA, 674–686. https://doi.org/10.1145/2382196.2382267

[29] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16).* ACM, New York, NY, USA, 622–633. https://doi.org/10.1145/2971648.2971760

[30] Jens Mattke, Lea Katharina Müller, and Christian Maier. 2017. Why do individuals block online ads? An explorative study to explain the use of ad blockers. In *23rd Americas Conference on Information Systems, AMCIS 2017, Boston, MA, USA, August 10-12, 2017.* Association for Information Systems.

[31] McAfee. 2017. McAfee WebAdvisor. (2017). Retrieved October 26, 2017 from https://home.mcafee.com/root/landingpage.aspx?lpname=get-it-now&affid=0&culture=en-sg

[32] Mozilla. 2017. Adblock Plus. (2017). Retrieved October 31, 2017 from https://addons.mozilla.org/en-US/firefox/addon/adblock-plus/

[33] Mozilla. 2017. Firefox Add-ons: Extensions. (2017). Retrieved October 31, 2017 from https://addons.mozilla.org/en-US/firefox/extensions/

[34] Mozilla. 2017. Firefox Add-ons: Most Popular Extensions. (2017). Retrieved October 31, 2017 from https://addons.mozilla.org/en-US/firefox/search/?sort=users&type=extension

[35] Mozilla. 2017. Mozilla Source Tree Docs: "main" ping. (2017). Retrieved October 23, 2017 from https://firefox-source-docs.mozilla.org/toolkit/components/telemetry/telemetry/data/main-ping.html

[36] Mozilla. 2017. Mozilla Source Tree Docs: Telemetry. (2017). Retrieved October 27, 2017 from https://firefox-source-docs.mozilla.org/toolkit/components/telemetry/telemetry/index.html

[37] Mozilla. 2017. uBlock Origin. (2017). Retrieved October 31, 2017 from https://addons.mozilla.org/en-US/firefox/addon/ublock-origin/

[38] Mozilla. 2017. Video Download Helper. (2017). Retrieved October 26, 2017 from https://addons.mozilla.org/en-US/firefox/addon/video-downloadhelper/

[39] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J Murdoch. 2016. Adblocking and Counter Blocking: A Slice of the Arms Race. In *6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16).* USENIX Association.

[40] Heather L. O'Brien and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology* 59, 6 (2008), 938–955. https://doi.org/10.1002/asi.20801

[41] Future of Privacy Forum. 2016. *Consumer Views Regarding Ad Blocking Technology.* Technical Report 65. Federal Trade Commission Public Comments: Consumer Privacy and Security Issues.

[42] PageFair. 2017. The state of the blocked web: 2017 Global Adblock Report. (2017). Retrieved October 26, 2017 from https://pagefair.com/downloads/2017/01/PageFair-2017-Adblock-Report.pdf

[43] Ofir Pele and Michael Werman. 2010. The Quadratic-Chi Histogram Distance Family. In *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 749–762.

[44] Eric T Peterson and Joseph Carrabis. 2008. *Measuring the immeasurable: Visitor engagement.* Technical Report. Web Analytics Demystified.

[45] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. 2015. Annoyed Users: Ads and Ad-Block Usage in the Wild. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15).* ACM, New York, NY, USA, 93–106. https://doi.org/10.1145/2815675.2815705

[46] Paul R. Rosenbaum. 2010. *Design of Observational Studies* (1 ed.). Springer-Verlag, New York, NY.

[47] Zahra Seyedghorban, Hossein Tahernejad, and Margaret Jekanyika Matanda. 2016. Reinquiry into advertising avoidance on the internet: A conceptual replication and extension. *Journal of Advertising* 45, 1 (2016), 120–129.

[48] Ben Shiller, Joel Waldfogel, and Johnny Ryan. 2017. *Will Ad Blocking Break the Internet?* Technical Report. National Bureau of Economic Research.

[49] Catherine E. Tucker. 2014. Social Networks, Personalized Advertising, and Privacy Controls. *Journal of Marketing Research* 51, 5 (2014), 546–562.