ABSTRACT

Consider two brands that want to jointly test alternate web experiences for their customers with an A/B test. Such collaborative tests are today enabled using third-party cookies, where each brand has information on the identity of visitors to another website, ensuring a consistent treatment experience. With the imminent elimination of third-party cookies, such A/B tests will become untenable. We propose a two-stage experimental design, where the two brands only need to agree on high-level aggregate parameters of the experiment to test the alternate experiences. Our design respects the privacy of customers. We propose an unbiased estimator of the Average Treatment Effect (ATE), and provide a way to use regression adjustment to improve this estimate. On real and simulated data, we show that the approach provides valid estimate of the ATE and is robust to the proportion of visitors overlapping across the brands. Our demonstration describes how a marketer can design such an experiment and analyze the results.

CCS CONCEPTS
• Applied computing → Marketing; Electronic commerce; • Security and privacy → Privacy protections;

KEYWORDS
Cookie-less internet, treatment effect, A/B testing, Advertising

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WSDM ’23, February 27–March 3, 2023, Singapore, Singapore
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9407-9/23/02...
$15.00
https://doi.org/10.1145/3539597.3573036

1 INTRODUCTION

A/B testing (randomized experiment) is the gold standard for optimizing consumer experiences on the web. Cookies have historically played an important role in ensuring that the same customer receives a consistent (or sticky) experience, thus ensuring the validity of the A/B test. Consider the A/B testing scenarios, where different brands collectively test alternate customer experiences. Some examples include brands under a hotel chain, different franchises under a sporting league, related clothing brands jointly owned by a holding company, or multiple departments of a government. Thus far, such A/B tests have relied on third-party cookies, which requires brand 1’s cookie to be available on brand 2’s website. This is the only mechanism to ensure that the same individual receives the same experience across both web properties.

A recent trend on the web has been the increased focus on customer privacy. This is embodied in laws like the General Data Protection Regulation (GDPR), which require explicit permission from the visitor for tracking web sessions. Another manifestation of this is that all major browser ecosystems are discontinuing third-party cookies. This presents significant challenges for brands that want to jointly optimize consumer experiences.

Formally, this is our question of interest. Consider a binary treatment deployed across two websites; for example, Brands 1 and 2 have jointly decided to optimize the user experience on their websites (e.g., the colour of the buy button or the discount offer). Let’s call the treatments 1 and 2. A user may visit one or both websites. In the ideal case, a user assigned to the treatment 1 group on the first website will also see treatment 1 on the second website. However, in the case where the user’s identity cannot be confirmed across the two websites (due to a lack of third-party cookies), the user is not guaranteed to see the same treatment across both websites. This leads to four potential treatment exposures, leading to the potential outcomes $Y_{11}$, $Y_{21}$, $Y_{12}$, and $Y_{22}$ (first and second positions denoting the treatments at the first and second websites). The potential outcomes of real interest to the brands are $Y_{11}$ and $Y_{22}. The question of interest is whether the treatment has any effect, which
in the world with cookies corresponds to the effect $E[Y_{11} - Y_{22}]$. This problem would be easily solved if the websites could share information about which users were exposed to which treatment. However, it is not clear whether one can estimate this effect without sharing such individual user-level information.

In our work, we present a multi-stage randomization design that allows the estimation of the desired effect without sharing any user-level information. Figure 1 describes our proposed design. The two brands only agree on two aggregate parameters of the test: (1) the two treatments being tested, (2) a notion of clusters (dividing the population into $C1$ and $C2$, e.g., geographies, time, device types), and (3) the nuisance parameter $\alpha (\neq 1/2)$. No individual or identity data needs to be shared between the brands, thus respecting customer privacy. First, we show that our proposed average treatment effect (ATE) is unbiased. Next, we further propose a way of performing regression adjustment, which further helps statistical power by using attributes of the web visitors. Finally, with experiments on both simulated and real-world web data, we show that our estimate has a lower bias than the naive ATE (difference of treatment means).

In our demo, we show how a marketer for a brand can use our design in a two-stage test and analyze the results.

2 RELATED WORK

The challenges of using cookie-level identifiers on the web as a proxy for the individual’s true identity is well studied [2]. But looking beyond cookies as the identity in digital marketing is less well studied [12]. The Privacy Sandbox [1] is an initiative with proposals tackling privacy related challenges for ad-targeting, delivery and measurement; but they are early proposals subject to changes.

Resolving the user’s identity can overcome most measurement and attribution issues, and much research has gone into stitching cookies and device identifiers [4, 7, 9]. But these strategies rely on using generic features such as IP to represent the same human user, or use more detailed but private data. The first persists privacy concerns and leads to inaccurate stitching, while the second relies on first party data from walled gardens and precludes any cross-ecosystem analysis. Businesses are also building their own unique identifiers based on people-based identifiers (Email ID, Device) and allow advertisers access to custom segments. However concerns about scalability and possible future regulation, means this is not a sustainable long term strategy. Media mix modeling is re-emerging as a complementary approach to user-level attribution [6].

The increase in the transient nature of the cookies leads to a significant increase in the fragmentation of web identities [2]. Treatment effect attenuation in presence of identity fragmentation and debiasing estimator for cookie-level estimates have been analysed [3, 8]. But both of these make very strong assumptions and target the loss of first-party cookies.

Our approach is instead based on conducting parallel experimentation followed by stratified aggregation. Instead of constructing user-level records, our approach works by running similar experiments within each strata and constructing group-level records (aggregation). Our final estimate of the treatment effect then combines the within strata treatment effects in a weighted manner. Our proposed estimation method is unbiased, but with a higher variance. However by relaxing the need for constructing user-level data, it retains privacy, works without further assumptions and can achieve a wider coverage. Our work is inspired by Hudgens and Halloran [5] which proposed a two-stage randomization procedure to account for interference. Our scenario however is different in that the same unit potentially receives multiple treatments instead of having exposure to treatments of other units. Furthermore, we do not make exposure level assumptions.

3 METHOD

Problem: Let’s say that our two websites have users in set $A$ and $B$. Note that $A \cap B \neq \emptyset$; so some users will visit both. When third party cookies are available, this set of users will get a consistent view (in terms of experience/offers/ads etc.), but without cookies this cannot be guaranteed, and users may receive multiple treatments. For each user $u_i \in A \cup B$ we have an associated feature vector $x_i$. Each user $u_i \in A$ can be exposed to treatments 1 or 2 on website 1. Similarly, each user $u_i \in B$ can be exposed to treatments 1 or 2 on website 2. Since the two brands collaborate without sharing data, our approach takes the perspective of one of the two brands (since they can only analyze visitors on their website), without loss of generality, let’s consider the first brand. If a user $u_i \in A \cap B$ then they have exposure to treatments from both websites. On the other hand those in $A \setminus B$ are only exposed to one website and we label their other exposure on the other website as “0”. Thus we have 6 potential outcome variables, i.e., $Y_{11}, Y_{22}, Y_{12}, Y_{21}, Y_{10}, Y_{20}$. Of these variables only one is observed for each user based on the treatment they received and whether they visited both websites. We want to estimate the ATE of shifting from treatment 2 to 1.

The challenges in estimation arises because of two main issues a) each website can only control allocation of treatments only to their users; and b) the identity of the shared users is unknown. Whenever a user $u_i$ visits, the websites can only access the features $x_i$; and choose a treatment without information on whether $u_i \in A \cup B$ or how the other website might allocate treatment to this individual. By $\tilde{Y}$, we denote the random variable corresponding to the observed outcome on the website, i.e., for a user of website 1 and allocated to treatment 1, we observe $\tilde{Y}_{11}$. Since we do not know the treatment on the other website, $\tilde{Y}_{11}$ can be $Y_{11}, Y_{12}, Y_{10}, Y_{21}, Y_{20}$. On the other hand, in standard causal inference we know for each observation unit, whether the observed outcome is $Y_{11}$ or $Y_{22}$.

For a group which has been allocated treatment 1 by website 1, the expected average outcome is given by:

$$E[\tilde{Y}_{11}] = (1 - p)E[Y_{11}] + p(\alpha E[Y_{11}] + (1 - \alpha)E[Y_{12}]).$$

Here, $p$ is the fraction of users who are shared and hence visit both websites, while $\alpha$ is the fraction of these shared users who receive treatment 1 on the second website as well. Every user has the probability $1 - p$ of only visiting website 1 and hence consistently receives treatment 1. For these users the average outcome is $E[Y_{11}]$. For the rest of the users (who are $p$ fraction of the population), an $\alpha$ fraction of them are allocated to treatment 1 by website 2 (and hence are exposed to treatment pair 1, 2). The observed outcome on these users is $E[Y_{11}, Y_{12}]$. Similarly $1 - \alpha$ fraction of the shared users receive treatment 2, corresponding to average outcome $E[Y_{12}]$.

2We make two standard assumptions in treatment effect literature, strong ignorability and positivity
By symmetry between the treatments, we can write a similar equation for the average outcome of group allocated treatment 2 by website 1. The naive estimator of ATE, which is the average outcome difference of treatment 1 over treatment 2, gives:

\[(1 - p)\mathbb{E}[Y_{1,0} - Y_{2,0}] + p(\alpha \mathbb{E}[Y_{1,1} - Y_{2,1}] + (1 - \alpha)\mathbb{E}[Y_{1,2} - Y_{2,2}])\]

Next, we analyze the treatment effect if one could track the users and provide them a consistent experience. Every user has the probability \(1 - p\) of only being on website 1 and hence receiving only treatment 1 or 2. For these users the treatment effect is \(\mathbb{E}[Y_{1,0} - Y_{2,0}]\). The rest of the users visits both websites and would either receive only one of treatment 1 or 2 on both sites. The corresponding treatment effect would hence be \(\mathbb{E}[Y_{1,1} - Y_{2,1}]\). The average treatment effect of the entire population is then given by:

\[TE = (1 - p)\mathbb{E}[Y_{1,0} - Y_{2,0}] + p(\mathbb{E}[Y_{1,1} - Y_{2,1}])\]

It is clear from the desired treatment effect and the observed treatment effect are mismatched due to contributions from the cross treatment outcomes \(Y_{2,1}\) and \(Y_{1,2}\) (above equation). Moreover it is clear that mismatch increases with the fraction of shared users \(p\).

**Method:** We assume that we are able to split the population into more than one macro-cluster, with freedom to determine treatment allocation within each. Each user’s cluster membership determines their treatment allocation strategy, performed by randomization. Such multi-stage randomization strategies are well known in literature [5]. One way to achieve this is to do stratified randomization with a common cluster definition on both websites. More specifically for each user \(u_i\), we determine their cluster based on their features \(x_i\), and this function or mapping to macro-clusters is shared across all websites. Note that privacy is maintained here; since no website shares user-specific information with the other.

For simplicity consider only two clusters \(C_1\) and \(C_2\). In the first cluster, the allocation ratio of the treatment is chosen by the second website as \(\alpha \neq 0.5\); while in the second cluster the allocation ratio is \(1 - \alpha\) (brand 1 uses the opposite allocation ratios) (see Figure 2 for a flowchart). We denote the empirical outcomes for cluster \(C_i\) and treatment \(j\) on brand 1 by the variable \(\bar{Y}^{C_j}_i\) (i.e., it is the sample estimate of \(\mathbb{E}(Y^{C_j}_j)\)). Note that only the details of randomization system are shared, not any individual or group means. Then given our setting we have the four observed outcomes \(\bar{Y}^{C_1}_1 + \bar{Y}^{C_2}_1, \bar{Y}^{C_1}_2, \bar{Y}^{C_2}_2\) (all for brand 1). One mechanism we first conduct simulation experiments where, by design, all parameters are known and adjustable. We can then quantitatively measure the performance of our method across ranges of key parameters. This simulation was conducted by generating 10,000 observations, and each experiment was repeated 20 times. Each potential outcome variable \(Y_{ij}\) at a single observation unit is obtained via a gaussian linear model from covariates \(X\). We vary two parameters \(\delta_1 = \mathbb{E}[Y_{1,1}] - \mathbb{E}[Y_{1,2}]\) and \(\delta_2 = \mathbb{E}[Y_{2,1}] - \mathbb{E}[Y_{2,2}]\).

Results We conducted these simulations and measured the error in the estimated ATE versus the true ATE for four methods. These include the standard ATE estimate (“uncorrected”), standard ATE with covariate adjustment (“uncorrected + adj”), our ATE estimate (“corrected”) and our ATE estimate with covariate adjustment.

![Figure 2: Flowchart of clustered treatment allocations for A/B testing. The user clustering can be as per Figure 1](image)

### Algorithm 1 Covariate Adjustment Algorithm

**Input:** Vector of outcomes \(Y_{1,1}^{C_1}, Y_{2,1}^{C_1}, Y_{1,2}^{C_1}, Y_{2,2}^{C_1}\), allocation \(\alpha\)

**Output:** Covariate adjusted treatment effect \(\hat{ATE}_{cov}\)

(1) Let \(Z_{ij}^{C_j}\) i.e. it is \(1\) if \(Y\) corresponds to treatment 1 and 0 otherwise

(2) Fit OLS(\(Y_{1,1}^{C_1}, Y_{2,1}^{C_1}\) ~ \(X + 1 + [Z_{1,1}^{C_1}, Z_{2,1}^{C_1}]\))

(3) Fit OLS(\(Y_{1,2}^{C_1}, Y_{2,2}^{C_1}\) ~ \(X + 1 + [Z_{1,2}^{C_1}, Z_{2,2}^{C_1}]\))

(4) Let \(\beta_1, \beta_2\) be coefficient of \(Z\) as estimated in steps 2.3

(5) \(\hat{ATE}_{cov} = \frac{1}{2\alpha - 1}(\alpha(\hat{Y}_{1,1} - \hat{Y}_{1,2}) + (1 - \alpha)(\hat{Y}_{2,1} - \hat{Y}_{1,2}))\)

---

This work is a shortened version of Shankar et al. [10]
We develop a demonstration of our technology that shows how we measured the error in the estimated ATE versus the true value with proportion of users (who is present in only one time period corresponds to a user who visits only one website). This splitting allows us to create a joint distribution of outcomes and treatments to sample from.

Scenario 1: We split the total time period into two halves, and considered the visitors in the two periods as visiting two separate websites. A user who is present in both time periods is a user with exposure to both parts of the treatment, while a user who is present in only one time period corresponds to a user who visits only one website. This splitting allows us to create a joint distribution of outcomes and treatments to sample from.

Scenario 2: We ran another experiment where we directly isolated users who were exposed to multiple treatments and estimated the ATE from this data. Since the outcomes in this case are sampled from historical logs, the effect of interference is fixed and cannot be changed. However one can analyze the effect of changing the fraction of users who receive exposure to multiple treatments.

Results: In both the scenarios, the underlying true exposures to both websites are known, allowing us access to the true ATE. We measured the error in the estimated ATE versus the true value and present the results in Figure 4. The plots depicts both the bias of the estimator and its standard error. In line with the theoretical analysis earlier, the bias of the standard estimator linearly increases with proportion of users (p) exposed to multiple treatments.

4.3 Demonstration

We develop a demonstration of our technology that shows how a marketer for a brand collaborating on conducting an A/B test jointly with another brand can design and analyse an experiment by only sharing the parameters of the experiment. In Figure 5 the marketer for Brand A specifies the allocation ration of \( \alpha \). The chart shows the estimated treatment effect estimated from the naive difference of treatment differences, and the proposed ATE estimate. Our approach suggests that there is no treatment effect (which is the truth in this simulated scenario), while the naive estimate would suggest a negative treatment effect. Similarly, a marketer for Brand B, can explore the same experiment from their perspective.

5 CONCLUSION

We have proposed a two-stage experimental design that brands can use to jointly estimate and test the treatment effect from an experiment. Our approach has no dependence on third-party cookies, and does not require any individual level information to be shared between the brands. We show that the proposed ATE estimate is unbiased, and show how to compute regression adjusted estimate of the ATE. Future work includes addressing multiple treatment arms and exploring effect of fragmentation on a single brand’s site.

REFERENCES

[1] Dieter Bohn. 2019. Building a more private web. (2019). https://bit.ly/2ZmoHyY
[2] Patrali Chatterjee, Donna Hoffman, and Thomas Novak. 2003. Modeling the clickstream: Implications for web-based advertising. Marketing Science (2003).
[3] Dominic Coey and Michael Bailey. 2016. People and cookies: Imperfect treatment assignment in online experiments. In Proceedings of 25th WWW.
[4] Johannes De Smedt, Ewelina Lacka, Spyro Nita, and Hans-Helmut Kohls. 2021. Session stitching using sequence fingerprinting for web page visits. (2021).
[5] Michael G. Hudgens and M. Elizabeth Halloran. 2008. Toward causal inference with interference. J. Amer. Statist. Assoc. 103, 482 (2008), 832–842.
[6] Roger Kamena. 2021. Moving towards inferential attribution modelling in a world without third-party cookies. Applied Marketing Analytics 7, 2 (2021), 122–130.
[7] Sungshul Kim, Nikhil Kini, Jay Pujara, Eunyee Koh, and Lise Getoor. 2017. Probabilistic visitor stitching on cross-device web logs. In Proceedings of 26th WWW.
[8] Tesary Lin and Sanjog Misra. 2021. The identity fragmentation bias. (2021).
[9] Eshwar Saha Roy, Ritwik Sinha, Niyati Chhaya, and Shiv Saini. 2015. Probabilistic deduplication of anonymous web traffic. In Proceedings of 26th WWW.
[10] Shiv Shankar, Ritwik Sinha, Saayan Mitra, Moumita Sinha, Vishy Swaminathan, and Sridhar Mahadevan. 2022. Privacy Aware Experiments without Cookies. OpenReview Preprint (2022). https://openreview.net/forum?id=-CIIHevXoho
[11] Ritwik Sinha, Ivan Andrus, and Trevor Paulsen. 2020. Attribution IQ: Scalable Game Theoretic Attribution in Web Analytics. In Proceedings of 29th ACM CIKM.
[12] Ian Thomas. 2021. Planning for a cookie-less future: How privacy changes will impact marketing, targeting and analytics. Applied Marketing Analytics 7 (2021).