People and Cookies: Imperfect Treatment Assignment in Online Experiments

Dominic Coey
Core Data Science, Facebook
1 Hacker Way
Menlo Park, CA 94025
coey@fb.com

Michael Bailey
Core Data Science, Facebook
1 Hacker Way
Menlo Park, CA 94025
mcbailey@fb.com

ABSTRACT

Identifying the same internet user across devices or over time is often infeasible. This presents a problem for online experiments, as it precludes person-level randomization. Randomization must instead be done using imperfect proxies for people, like cookies, email addresses, or device identifiers. Users may be partially treated and partially untreated as some of their cookies are assigned to the test group and some to the control group, complicating statistical inference. We show that the estimated treatment effect in a cookie-level experiment converges to a weighted average of the marginal effects of treating more of a user's cookies. If the marginal effects of cookie treatment exposure are positive and constant, it underestimates the true person-level effect by a factor equal to the number of cookies per person. Using two separate datasets—cookie assignment data from Atlas and advertising exposure and purchase data from Facebook—we empirically quantify the differences between cookie and person-level advertising effectiveness experiments. The effects are substantial: cookie tests underestimate the true person-level effects by a factor of about three, and require two to three times the number of people to achieve the same power as a test with perfect treatment assignment.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Experimental design*

General Terms

Economics

Keywords

Advertising effectiveness, causal inference, cookies, experiments, online advertising

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2882984 .

1. INTRODUCTION

One major limitation of online experiments is that the experimenter often does not have complete control over who is exposed to the treatment [7, 8, 9, 28]. In contrast to experiments conducted in person, in which it is straightforward to assign non-overlapping groups of people to the test and control groups, the online experimenter often does not have the ability to identify the same person across devices or over time. She must instead resort to randomizing using a proxy which imperfectly identifies users, like a cookie, an email address, or an account or device identifier.

Cookies are the most common technology used to identify online users and devices making cookie-based experiments especially popular among researchers and technology companies [11, 21, 25, 28]. A *cookie* is a small piece of data sent from the website and stored on the user's browser that is sent back to the website every time the user returns. The same user may generate multiple cookies by clearing his cookies and being assigned new ones (*cookie churn*), using multiple browsers, or visiting the same website on different devices. Additionally, some browsers will delete cookies on crashing, remove older cookies, and cookies can become corrupted leading to the same user using the same browser being assigned different cookies. If treatment assignment is randomized at the cookie level, the same user may sometimes be assigned to the test group and sometimes to the control group, depending on when he visits the website and the browser or device he is using. People with cookies in the test group are only partially treated, and those with cookies in the control group are only partially untreated. This complication makes it unclear what information the comparison of test and control cookie outcomes provides.

The same problem arises with experiments using other proxies like email addresses, device identifiers, or user accounts, as a single user may be assigned to different proxies and thus different conditions in the experiment. In this paper we focus on cookies since it is the most widely used identity technology on the internet, especially for advertisers who want to distinguish customers, but our analysis and results apply generally to any proxy that doesn't have perfect assignment to users.

We show that the test-control cookie comparison estimates a weighted average of the marginal effects on a user of having an additional cookie exposed to the treatment. In contrast to the ideal experiment in which users can be perfectly assigned to test or control, this weighted average depends on the probability that a cookie is assigned to the test group. Changing the test group assignment probabil-

ity changes the quantity estimated. Failing to replicate the results of an experiment which used a different assignment probability does not necessarily indicate that the initial results are invalid.

If the marginal treatment effects are all positive, i.e. treating more of the user's cookies always increases the mean of the outcome variable, then the test-control cookie comparison underestimates the person-level treatment effect in which users are randomized into test and control groups. This provides a formal justification for the folk wisdom that cookie-based experiments tend to attenuate the true treatment effects. If in addition to being positive the marginal treatment effects are constant (or if they are affine and cookies are assigned to the test group with probability 0.5), the person-level treatment effect is greater than the test-control cookie comparison by a factor equal to the number of cookies per user. This result is intuitive: as the number of cookies increases the average difference in outcomes between test and control cookies becomes smaller.

We use a unique dataset to quantify how much imperfect treatment assignment matters in practice in the context of measuring advertising effectiveness. *Atlas*[1] by Facebook allows advertisers to serve ads across third-party websites and mobile apps. Atlas cookies contain a one-way hashed version of the individual's Facebook identifier for Facebook users. Because we observe both the Atlas cookie assigned to the user at the time of the ad impression as well as the hashed Facebook identifier for Facebook users, we have ground truth data on cookie assignment distributions.

We also use data from Facebook's *Conversion Lift*[2] product, which allows advertisers to run advertising effectiveness experiments. Facebook assigns a randomly selected group of users to a control group, which is not exposed to an advertising campaign, and compares their outcomes to the test group, which is eligible to see the campaign. By comparing online sales outcomes between test and control users, advertisers can determine how effective the advertisement is in increasing sales.

These two datasets–Atlas data on cookie assignments, and Conversion Lift data on ad exposure and sales–together enable us to simulate the effect of imperfect treatment assignment in ad effectiveness studies. The effects estimated in ad experiments with imperfect treatment assignment are rarely of inherent interest. Rather, the real effects of interest are typically the effects of fully rolling out the ad campaign vs. not advertising, as knowing those effects enables advertisers to determine if their ads are giving a sufficient return on investment. Equivalently, they are the effects that would be estimated by an experiment with perfect treatment assignment.

We find that cookie-based tests underestimate these person-level effects by a factor of around three. In addition, to achieve the same level of statistical power in a cookie-level experiment as a person-level experiment, around two to three times greater sample sizes are needed. Difficulties in detecting statistically significant effects in online experiments may be due to imperfect treatment assignment, rather than the true underlying effect sizes being small.

This paper complements the econometrics literature on instrumental variables and imperfect treatment assignment [1, 2, 4, 13, 14, 16]. The literature focuses on the case where the experimenter is only partially able to control individuals' treatment status. A physician may prescribe a drug, for example, but is unable to force his patient to take it. Our paper differs from this setting in that the experimenter can fully control individuals' treatment status for any of their cookies, but the fact that people have multiple cookies means they may end up being only partially treated or untreated. To the best of our knowledge, we are the first to formally treat and empirically analyze the problem of imperfect treatment assignment due to difficulties in identifying people online, rather than difficulties in influencing their behavior.

Our paper is also closely related to the growing body of work on online advertising effectiveness. A large number of prior studies [10, 15, 17, 23, 25, 27, 30] run advertising field experiments at the cookie level, and our work quantifies the potential measurement error in these studies from multiple cookie assignments. Other studies run advertising field experiments at the person level [26, 18, 3, 24], and our results indicate that treatment effects should not be directly compared between cookie and person-level experiments. Lewis et al. [24] find that online advertising campaigns often require relatively large samples to detect a significant effect on sales, highlighting the need to analyze the loss in statistical power from cookie assignments.

Because it is so rare to find data on cookie assignments by person, we hope to provide researchers with some context for understanding the extent of the bias in their cookie-level studies, if they believe the users in their study are similar to the population of U.S. Facebook users.

Finally, this work contributes to the rapidly growing literature on the challenges of implementing experiments online [6, 20, 22, 24]. There are several examples of how causal inference can be biased in online experiments including, among others, interference between test and control groups [5], correlated behaviors biasing observational studies [25], and "carryover" effects [20]. We demonstrate another major implementation challenge in that treatment effects are substantially attenuated when comparing cookie-level outcomes.

2. MODEL SETUP

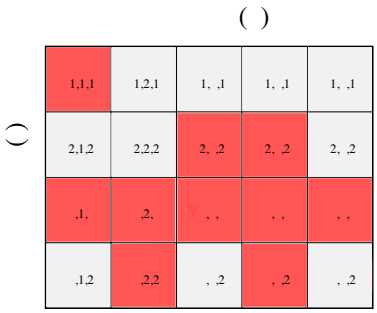
Each person $i \in \{1, \ldots, m\}$ generates n cookies. Each of i's cookies is independently assigned to be treated with probability p. Person i's outcome associated with cookie k if he were to have e treated cookies in total is a random variable denoted $y_{i,k,e}$. Across k the $y_{i,k,e}$'s are distributed with mean $\mu_i(e)$, so that $\mu_i(e)$ is i's expected cookie-level outcome when e of his cookies have been treated. Define the random variable $T_{i,k}$ where $T_{i,k}$ = 1 if i's k^{th} cookie has been selected for treatment, and $T_{i,k}$ = 0 otherwise. We assume that $y_{i,k,e}$ and $T_{i,k}$ are independent: given i's number of treated cookies, his treatment and control cookies' outcomes are the same on average.

Let $e_i = \sum_{k=1}^{n} T_{i,k}$ denote the number of treatments that i receives. For each of i's cookies, the researcher observes the outcome variables $y_{i,1,e_i}, y_{i,2,e_i}, \ldots, y_{i,n,e_i}$ associated with the n cookies. For example, in an advertising experiment run by an online retailer, the outcome variable of interest, y_{i,k,e_i}, might be the amount of spending attributable to user i's k^{th} cookie, given that user i has seen e_i advertisements across his n devices, each of which has a different cookie.

[1] http://atlassolutions.com/

[2] https://www.facebook.com/business/news/conversion-lift-measurement

Figure 1: Treatment Assignment



Random cookie assignment to treatment and control groups. Each row of cookies belongs to the same person. Red squares are treated cookies, gray squares are untreated cookies.

The *cookie-level treatment e□ect estimator*, $\widehat{C}$, is the average over treatment cookies of their associated outcomes, minus the corresponding average over control cookies:

$$\widehat{C} \equiv \frac{\sum_{i=1}^{m}\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i}}{\sum_{i=1}^{m} e_i} - \frac{\sum_{i=1}^{m}\sum_{k=1}^{n}(1-T_{i,k}) y_{i,k,e_i}}{\sum_{i=1}^{m}(n-e_i)}.$$

Although the researcher can calculate the sum of outcomes associated with all test cookies, $\sum_{i=1}^{m}\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i}$, she cannot calculate the summands $\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i}$ for any i (and similarly for control outcomes). This is because she has no way of identifying which cookies or outcomes are associated with the same person. Figure 1 shows a simple example where $m = 4$ and $n = 5$. In terms of the figure, $\widehat{C}$ is the average of the values in red squares minus the average of the values in gray squares. The researcher observes whether an outcome is red or gray, but not whether two outcomes belong to the same row. What does the cookie-level treatment e□ect estimator estimate, and how does it relate to the expected marginal e□ects $E[\mu_i(j+1) - \mu_i(j)]$, or to the e□ect of fully treating users, $E[\mu_i(n) - \mu_i(0)]$?

3. MODEL ANALYSIS

We begin by proving that the cookie-level estimator $\widehat{C}$ converges to a weighted average of the expected marginal e□ects $E[\mu_i(j+1) - \mu_i(j)]$, where the weights are the probabilities of a binomial distribution.

PROPOSITION 1. *Let X be a random variable with distribution $B(n-1, p)$, independent of the μ_i, and assume the expectations $E[\mu_i(e)]$ are finite for all e. Then $\widehat{C} \to_p E[\mu_i(X+1) - \mu_i(X)]$.*

PROOF. By the weak law of large numbers and the continuous mapping theorem, we have

$$\frac{m^{-1}\sum_{i=1}^{m}\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i}}{m^{-1}\sum_{i=1}^{m} e_i} - \frac{m^{-1}\sum_{i=1}^{m}\sum_{k=1}^{n}(1-T_{i,k}) y_{i,k,e_i}}{m^{-1}\sum_{i=1}^{m}(n-e_i)}$$
$$\to_p \frac{E[\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i}]}{E[e_i]} - \frac{E[\sum_{k=1}^{n}(1-T_{i,k}) y_{i,k,e_i}]}{E[n-e_i]}.$$

Expectations in expressions involving people (indexed by i) and cookies (indexed by k) are over both variables. The probability limit on the right hand side can be rewritten as

$$\frac{\sum_{j=0}^{n} P(e_i = j) E[\sum_{k=1}^{n} T_{i,k} y_{i,k,e_i} | e_i = j]}{\sum_{j=0}^{n} P(e_i = j) j} - \frac{\sum_{j=0}^{n} P(e_i = j) E[\sum_{k=1}^{n}(1-T_{i,k}) y_{i,k,e_i} | e_i = j]}{\sum_{j=0}^{n} P(e_i = j)(n-j)}$$
$$= \frac{\sum_{j=0}^{n} P(e_i = j) j E[\mu_i(j)]}{\sum_{j=0}^{n} P(e_i = j) j} - \frac{\sum_{j=0}^{n} P(e_i = j)(n-j) E[\mu_i(j)]}{\sum_{j=0}^{n} P(e_i = j)(n-j)},$$

where the first step iterates expectations over e_i, and the second uses independence of $y_{i,k,e}$ and $T_{i,k}$, as well as the fact that $e_i = \sum_{k=1}^{n} T_{i,k}$.

Given that $e_i \sim B(n, p)$, the last expression can be rewritten as follows:

$$\frac{\sum_{j=0}^{n} P(e_i = j) j E[\mu_i(j)]}{\sum_{j=0}^{n} P(e_i = j) j} - \frac{\sum_{j=0}^{n} P(e_i = j)(n-j) E[\mu_i(j)]}{\sum_{j=0}^{n} P(e_i = j)(n-j)}$$
$$= \sum_{j=0}^{n} \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j} \frac{j}{np} E[\mu_i(j)]$$
$$- \sum_{j=0}^{n} \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j} \frac{n-j}{n(1-p)} E[\mu_i(j)]$$
$$= \sum_{j=1}^{n} \frac{(n-1)!}{(j-1)!(n-j)!} p^{j-1} (1-p)^{n-j} E[\mu_i(j)]$$
$$- \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1} E[\mu_i(j)]$$
$$= \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1} E[\mu_i(j+1)]$$
$$- \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1} E[\mu_i(j)]$$
$$= \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1} (E[\mu_i(j+1)] - E[\mu_i(j)])$$
$$= E[\mu_i(X+1) - \mu_i(X)],$$

where $X \sim B(n-1, p)$. □

This result shows that the cookie-based measure estimates a weighted average of all the marginal e□ects of an extra treatment exposure, where the weights are given by the probability mass function of a $B(n-1, p)$ random variable. More weight is placed on the marginal e□ects at the likely levels of treatment exposure. This is rather intuitive–if p were close to one, then most of the test and control cookies would belong to people who have been treated a large number of times, and the experiment can not be very informative about the marginal e□ects of the first few exposures. The next corollary is immediate from Proposition 1.

COROLLARY 1. *If there is a constant marginal effect of receiving a treatment cookie instead of a control cookie (i.e. if $E[\mu_i(j+1)]-E[\mu_i(j)]$ does not depend on j), the cookie-based measure converges in probability to this marginal effect.*

Different treatment probabilities p result in different weightings of the marginal effects, and therefore estimate different quantities. Unless $E[\mu_i(j)]$ is affine, there is no reason to expect results from otherwise identical tests with different treatment probabilities to coincide, even ignoring sampling error. We denote $E[\mu_i(X+1)-\mu_i(X)]$, the probability limit of the cookie-based estimate $\widehat{C}$, as $C(p)$, to make dependence on p explicit. We recall a definition and a result on stochastic orderings of random variables, and prove that if $E[\mu_i(j)]$ is concave (convex), then $C(p)$ is decreasing (increasing) in p.

Definition 1. For random variables X and X' with cumulative distribution functions F_X and $F_{X'}$, X is said to *first-order stochastically dominate* X' if $F_X(c) \leq F_{X'}(c)$ for all c.

We write $X \geq_{FOSD} X'$ if X first-order stochastically dominates X'. The following classic result from [12] relates first-order stochastic dominance to means.

PROPOSITION 2. ([12]). *$X \geq_{FOSD} X'$ if and only if $Eu(X) \geq Eu(X')$ for all increasing functions u.*

This result allows us to derive the next proposition, which shows how the cookie-level effect varies with the cookie treatment probability.

PROPOSITION 3. *If $E[\mu_i(j)]$ is concave (convex) in j, then $C(p)$ is decreasing (increasing) in p.*

PROOF. If $X \sim B(n,p)$ and $X' \sim B(n,p')$ with $p \geq p'$, then $X \geq_{FOSD} X'$ ([19]). Set $u(j) = E[\mu_i(j+1)]-E[\mu_i(j)]$. If $E[\mu_i(j)]$ is concave is in j, then $u(\cdot)$ is decreasing. Proposition 2 implies that if $X \geq_{FOSD} X'$ and $u(\cdot)$ is decreasing, then $E[u(X)] \leq E[u(X')]$. It follows that $E[u(X)] \leq E[u(X')]$, i.e. that $C(p)$ is decreasing in p. The argument for convex $E[\mu_i(j)]$ is analogous. □

The effect of interest to the experimenter is typically the expected difference in outcomes when all cookies are treated and when no cookies are treated: $E[\mu_i(n) - \mu_i(0)]$. The reason this effect is important is that it is the causal effect of fully rolling-out the treatment (i.e. the causal effect of treating everyone all of the time, relative to never treating anyone), and the purpose of the experiment is often to decide whether the treatment should be rolled-out to everyone. It is also the effect that would be estimated in a user-level experiment: if people could be perfectly assigned to test and control, the average outcome would be $E[\mu_i(n)]$ in the test group and $E[\mu_i(0)]$ in the control group. This cookie-based effect $C(p)$ underestimates $E[\mu_i(n) - \mu_i(0)]$, if $E[\mu_i(j)]$ is increasing.

PROPOSITION 4. *If $E[\mu_i(j)]$ is increasing, then for all p, $C(p) \leq E[\mu_i(n) - \mu_i(0)]$.*

PROOF.

$$\begin{aligned} C(p) &= \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j (1-p)^{n-j-1} (E[\mu_i(j+1)] - E[\mu_i(j)]) \\ &\leq \sum_{j=0}^{n-1} (E[\mu_i(j+1)] - E[\mu_i(j)]) \\ &= E[\mu_i(n) - \mu_i(0)]. \end{aligned}$$

□

By how much does the cookie-based estimator underestimate the true effect of interest? It follows from Corollary 1 that if $E[\mu_i(j)]$ is affine in j, then $nC(p)$ is equal to the full rollout effect, $E[\mu_i(n) - \mu_i(0)]$, so that the cookie-based estimator is too small by a factor of n. If $E[\mu_i(j)]$ is not affine in j, the underestimation may be more or less severe, depending on the shape of the function $E[\mu_i(j)]$ as well as the cookie treatment probability p. We show that if the outcome response is quadratic and $p = 0.5$, then $nC(p)$ equals $E[\mu_i(n) - \mu_i(0)]$, just as in the affine case. More generally, whether $nC(p)$ overestimates or underestimates $E[\mu_i(n) - \mu_i(0)]$ depends not on whether outcomes themselves are concave or convex in the number of treatments received, but whether the *marginal* effect of an extra treatment on outcomes is concave or convex in the number of treatments received. Some further preliminaries on stochastic dominance are required.

Definition 2. For random variables X and X' with cumulative distribution functions F_X and $F_{X'}$, X is said to *second-order stochastically dominate* X' if $\int_{-\infty}^{c} [F_{X'}(c) - F_X(c)]dt \geq 0$ for all c.

We write $X \geq_{SOSD} X'$ if X second-order stochastically dominates X'. The next result, from [29], relates second-order stochastic dominance to means of concave transformations.

PROPOSITION 5. ([29]). *For random variables X and X' with the same mean, $X \geq_{SOSD} X'$ if and only if $Eu(X) \geq Eu(X')$ for all bounded concave functions u.*

This proposition allows us to derive some results about the relative sizes of the cookie-based estimator and the true user-level effect.

PROPOSITION 6. *If $p = 0.5$ and the marginal effects, $E[\mu_i(j+1)] - E[\mu_i(j)]$ are: i) concave in j, then $nC(p) \geq E[\mu_i(n) - \mu_i(0)]$; ii) convex in j, then $nC(p) \leq E[\mu_i(n) - \mu_i(0)]$.*

PROOF. i) Let $X \sim B(n-1, 0.5)$, let X' be uniformly distributed on $\{0, \ldots, n-1\}$, and define $u(j) = E[\mu_i(j+1)] - E[\mu_i(j)]$. The random variables X and X' have the same mean. X' can be obtained from X by a sequence of mean-preserving spreads, so $X \geq_{SOSD} X'$ (see [29]). By

assumption, the function u is concave. We have

$$\begin{aligned} nC(p) &= n(E[\mu_i(X+1) - \mu_i(X)]) \\ &= nE[u(X)] \\ &\geq nE[u(X')] \\ &= n\sum_{j=0}^{n-1} \frac{1}{n}(E[\mu_i(j+1)] - E[\mu_i(j)]) \\ &= E[\mu_i(n) - \mu_i(0)]. \end{aligned}$$

The first equality holds by the definition of $C(p)$ and the second by the definition of $u(X)$. The inequality follows from Proposition 5. The third equality follows by the definition of X'. Part ii) is analogous. □

To see the logic behind this proposition, note that $\frac{1}{n}E[\mu_i(n) - \mu_i(0)]$ is the unweighted average of all the marginal effects, whereas $C(p)$ is a weighted average of all the marginal effects, with relatively less weight on the extremes (that is, less weight on $E[\mu_i(j+1) - \mu_i(j)]$ for j close to 0 or n). The effect of shifting weight from the extreme to the intermediate marginal effects depends on the concavity or convexity of the sequence of marginal effects.[3]

An implication of this result is that when the outcome response is roughly quadratic, so that the marginal effects are roughly linear, the cookie treatment probability p should, if possible, be set to 0.5. This allows the true effect of fully rolling out the treatment to be captured by scaling the cookie-level treatment effect estimator by the number of cookies per person, while lower or higher p's could lead to potentially misleading estimates of the effect.

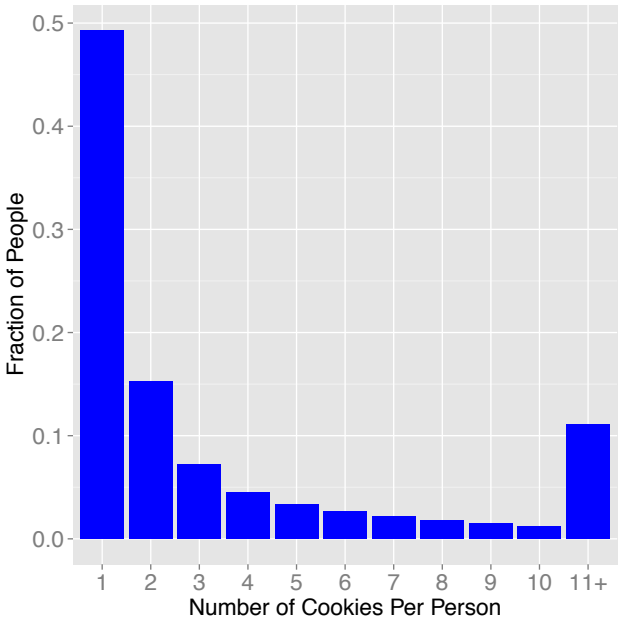
4. DATA AND EXPERIMENTAL SIMULATIONS

Proposition 4 shows that treatment-control differences are smaller on average in cookie-level experiments than people-level experiments. This raises two concerns. First, if cookie-based test estimates are incorrectly interpreted as reliable estimates of the true effect of rolling out a treatment, some treatments that are in fact worthwhile may not be implemented. Even when the attenuation bias in cookie-based estimates is understood, the uncertainty over the sizes of the true effects will hinder decision-making. Second, statistical power will likely suffer. A non-zero treatment effect may be less likely to be detected in a cookie-level experiment than a people-level experiment. The extent to which these issues are problems in practice depends on the distribution of the number of cookies per person. In the special case where each person is assigned a single cookie, for example, cookie and people-based tests are identical.

Obtaining data on the empirical distribution of cookies per person is generally difficult. It requires being able to match people to their cookies, and if this were straightforward there would be no need for cookie-based tests in the first place. Facebook's Atlas offering allows advertisers to serve ads on third-party websites and mobile applications. Atlas can group together the cookies associated with a Facebook user using a hashed Facebook id, when the user signs into their Facebook account. We thus observe both cookie assignments at the user-level for a group of users exposed to Atlas advertising campaigns. This gives us the ability to match cookies to people, across desktop and mobile devices, or different browsers, or over time, and calculate the distribution of cookies per person. This matching will be imperfect, as some users may never sign into Facebook on some of their devices. To the extent that we are underestimating the true number of cookies per person for this reason, the treatment dilution effects in our simulations are likely to understate the true effects.

The number of cookies per person observed in Atlas data over a one-month period (July 2015) is depicted in Figure 2. Slightly over half of people are observed to have more than one cookie during this period and over 10% are associated with over 10 cookies. In our simulations and throughout what follows, we deal with outliers by winzorising the data at 11 cookies, so that people who have over 11 cookies are treated as having exactly 11 cookies.[4]

Figure 2: Cookies Per Person, Histogram



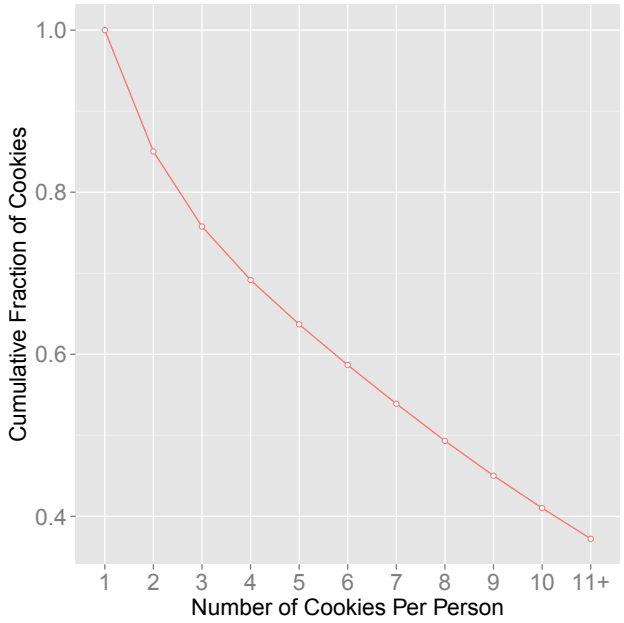
Even more directly related to treatment dilution effects is Figure 3, which shows the fraction of cookies belonging to people with at least a given number of cookies. The chief reason for treatment dilution is that cookies belonging to people with many cookies are relatively unhelpful in detecting an effect of the treatment, as those people have likely been exposed quite evenly to both treatment and control, and their treatment and control cookies are unlikely to be very different. In the limit, these peoples' cookies do nothing but add noise to the treatment and control comparison,

[3] Another classic result on stochastic dominance closely related to Proposition 5 is that $X \geq_{SOSD} X'$ if and only if $E[u(X)] \geq E[u(X')]$ for all nondecreasing and bounded concave functions u (see [12]). In a manner analogous to Proposition 6, this implies that if $p > 0.5$ and the marginal effects are nondecreasing and concave, then $nC(p) \geq E[\mu_i(n) - \mu_i(0)]$. Similarly if $p > 0.5$ and the marginal effects are nonincreasing and convex, then $nC(p) \leq E[\mu_i(n) - \mu_i(0)]$.

[4] This is conservative in the sense that it will tend to understate the true treatment dilution and loss of power associated with cookie-based testing.

and make it harder to detect a treatment effect. Figure 3 shows that these low-value cookies are frequent: about half of cookies come from people who have eight or more cookies. The single-cookie people, despite making up close to half of the population, only contribute about 15% of the cookies.

Figure 3: Fraction of Cookies Belonging To People With At Least n Cookies



Advertising effectiveness studies are a natural context for studying the difference between cookie and people-level experiments. They are particularly common uses of cookie-based testing [26, 27, 25], and also prone to lacking statistical power [24]. We use data from Facebook's Conversion Lift product which allows advertisers to run ad effectiveness experiments by designating a random subset of users to be in the hold-out or control group, and who will not see the ads. By comparing outcome data between the test and control group, advertisers can estimate the effectiveness of the advertisement in driving conversions. Advertisers will specify objectives for their campaign which typically are either continuous outcomes (e.g., total online sales) or binary outcomes (e.g., sign-ups, application installs).

For each Conversion Lift campaign we observe assignment to test or control for each user id, determining eligibility to see the campaign, and online outcomes generated on the advertiser side. These outcomes occur on the advertisers' website (e.g., purchasing a product from their online store) and the advertiser installs a conversion pixel that fires when the user takes the appropriate action, sending the outcome data back to Facebook to be matched back to the user's treatment status. Using the aggregate distributions from each dataset—user ids and cookie assignments from Atlas, and outcomes for test and control users from Conversion Lift studies—we can simulate the effect of running an advertising effectiveness experiment at both the cookie and person level.

We select two campaigns from the Conversion Lift program for our simulations. The first is aimed at generating user sign-ups for a online product, and the objective of the second is to increase online sales. This allows us to assess how cookie-based tests perform, for both Bernoulli and continuous outcome distributions (sign-ups and spending). In both campaigns the test and control groups are significantly different, albeit to quite different extents (with t-statistics of 5.47 in the sign-up campaign and 2.20 in the user campaign), making them good candidates for studying how statistically detectable effects are attenuated in cookie-based experiments.

For the cookie-level test simulation, we simulate the outcomes of m people, each of which has a total number of cookies n drawn independently across people from the empirical distribution depicted in Figure 2. Each cookie is independently assigned to the treatment group with probability 0.5.[5] The person-level outcomes of interest are independent across people, and denoted by the random variable $Y_{j,n}$, where j is the number of treatment cookies a person is exposed to, and n is his total number of cookies.[6] This person-level outcome is assigned uniformly at random to one of that person's n cookies.[7]

The random variables $Y_{j,n}$ are determined by the actual advertising effectiveness data. In both the user sign-up and spending simulations, the random variable $Y_{j,n}$ is defined as a mixture distribution: with probability j/n we draw from the empirical distribution of the corresponding test outcomes, and otherwise we draw from the empirical distribution of control outcomes. Thus the more "treated" a user is, the more likely he is to have an outcome drawn from the test distribution.

In our other simulations the linearity of the treatment effect, corresponding here to the mixture probability j/n being linear in j, appears not to be a critical determinant of the magnitude of treatment dilution or statistical power. This is unsurprising—by Proposition 1, treatment response curves with very different degrees of curvature will generate similar cookie-level treatment effects, as long as their weighted marginal effects are similar.

The person-level simulation is identical, except each person will either have all of his cookies treated or all of his cookies untreated, with each outcome being equally likely. Treated people will draw outcomes from the treatment distribution with certainty, and otherwise will draw outcomes from the control distribution with certainty.

This simulation procedure allows us to describe quantitatively the treatment dilution from cookie-based tests. Given the empirical distribution of cookies per person, the effect estimated in a cookie-based test is 30.4% of the people-based test effect for both the user sign-up advertising campaign, and the spending advertising campaign. This ratio depends only on the distribution of the number of cookies per person and not on the outcome distributions, as the ratio of the marginal effect of treating an extra cookie to the effect of

[5] In practice most advertisers use an unbalanced treatment assignment of 0.95, giving less power in both cookie and people-level tests.

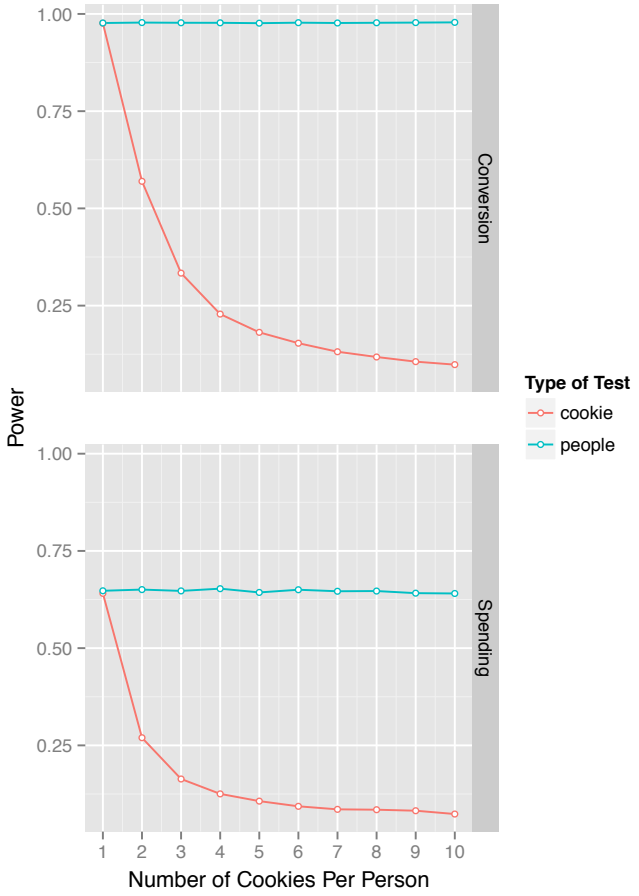
[6] In terms of the model of Section 2, for person i, $Y_{i,j,n}$ is the sum of the per-cookie outcomes: $Y_{i,j,n} = \sum_{k=1}^{n} T_{i,k} y_{i,k,j}$.

[7] The assumption that a single cookie is assigned the entire person-level outcome is a reasonable approximation in our data. In the sign-up experiment, it necessarily holds, as users cannot create duplicate accounts. In the spending experiment, the total number of transactions is just 4% higher than the total number of buyers, implying that relatively few buyers are transacting multiple times on different cookies.

treating all cookies is the same for both outcome distributions in this simulation.[8] Treatments that appear to give a positive return on investment on the basis of a cookie-based test can in fact be substantially more beneficial than the cookie test suggests.

To calculate statistical power, we repeatedly conduct the simulations described above and for each simulation calculate the t-statistic associated with the null hypothesis of no difference in means between test and control outcomes, using the standard form of the t-test for groups with unequal variances. We reject at a 5% one-sided level. We draw outcomes for all people 10,000 times, producing 10,000 t-statistics. The test's power is calculated as the fraction of simulations for which the t-statistic lies in the rejection region (i.e. above 1.64).[9]

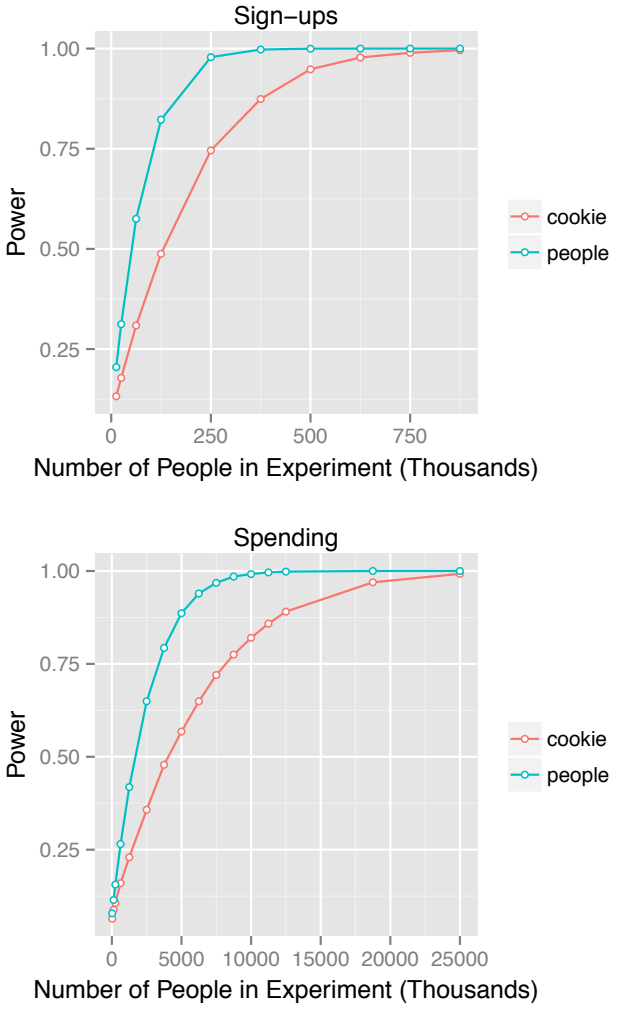
Figure 4: Power And Number of Cookies Per Person



Before using the Atlas cookie data, it is useful to quantify exactly how much test precision is affected by the number of cookies per person. Figure 4 shows how statistical power decreases as a function of the number of cookies per person, with 250,000 people in the sign-up experiment and 2.5m people in the spending experiment. With one cookie per person, the people and cookie tests are equivalent. Power in the cookie tests declines sharply as the number of cookies per person increases, while it remains constant in the people tests. With two cookies, cookie-based test power drops by 41% in the sign-up experiment and 59% in the spending experiment. With five or more cookies per person, cookie-based tests are so underpowered as to be of rather limited value.

Next, we incorporate the data on the actual distribution of cookies per user. Figure 5 shows how the power of cookie- and people-based test compare for different sample sizes, and for the two outcome distributions. Both tests are consistent, in that the null hypothesis will be rejected with probability approaching one as sample sizes increase. However for a fixed sample size, cookie-based tests are considerably less powerful than people-based tests. With 2.5 million people in the spending experiment, for example, the null hypothesis will be correctly rejected in 65% of people-level experiments, but only 36% of cookie-level experiments. With 125,000 people in the sign-up experiment, the corresponding figures are 82% and 49%.

Figure 5: Power And Sample Size

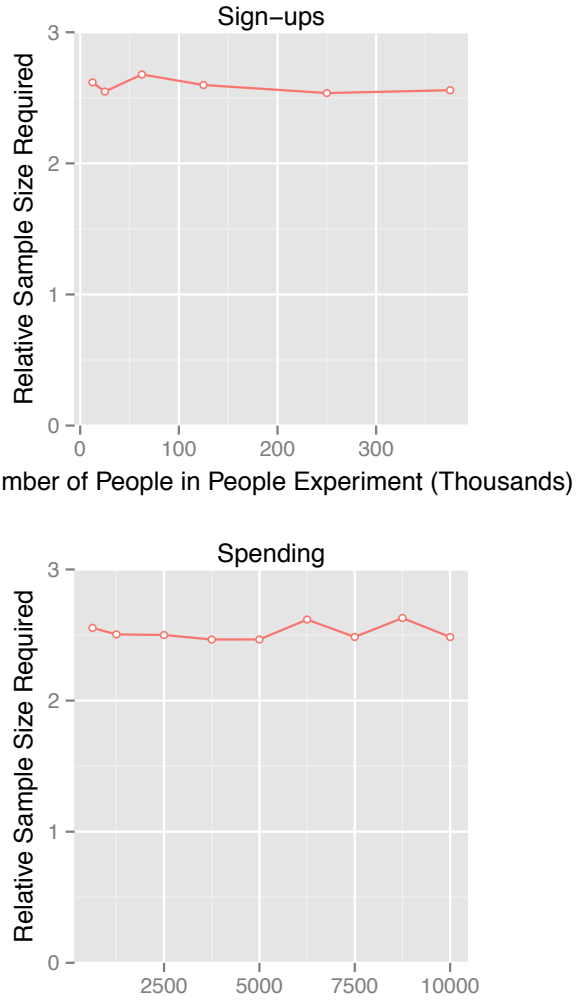


[8] We use m = 1 billion people to estimate mean spending for the treated and untreated cookies and people.

[9] The t-statistics are slightly different in the people and cookie cases, as the relevant number of observations to be used in constructing the statistics should be either the number of people or number of cookies, as appropriate.

Figure 6 gives a di□erent perspective on statistical power. It shows the factor by which the number of people in a cookie-level test must exceed the number of people in a people-level test, to achieve the same power. Equivalently, the figure describes how much larger sample sizes must be to make up for the precision lost in cookie testing. These relative sample sizes are shown as a function of sample size in the people-level test. Overall, cookie tests need to have 2 to 3 times as many people as people tests to achieve the same precision.[10]

Figure 6: Relative Sample Sizes Required, Cookies vs. People



One implication of Figure 6 is that the data on the 51% of people with more than one cookie, and the 85% of cookies they contribute, are of *negative* value in these simulations. Under the assumptions of the simulation, where treatment e□ects do not vary by the total number of cookies a person is assigned, the experimenter would be better o□ having no data at all on these people. In a 200,000 person cookie-level test, about 100,000 people have a single cookie. If it were possible to restrict attention just to this group's data, this would be a people-level test with 100,000 people. From the results of Figure 6, this is equal in power to a cookie-level test with over 200,000 people, and so more powerful than the initial sample size of 200,000.

[10]We compute cookie-level power as a function of sample size, as required for this calculation, by linearly interpolating through the points in Figure 5. For sample sizes beyond about 400,000 people in the sign-up simulation, the numerical imprecision of this interpolation becomes more substantial as statistical power asymptotes to 1.

5. CONCLUSION

Our theory and simulations calibrated with actual advertising e□ectiveness and cookie data suggest that imperfect treatment assignment can substantially reduce the di□erences in average outcomes between test and control groups, and may present a serious obstacle to learning about the true underlying treatment e□ects in online experiments. Cookie-based testing is likely to introduce a status-quo bias into decision-making, both because it reduces the probability of finding significant e□ects, and because it attentuates the estimatedenefits of the treatmenteing tested.

Although the level of randomization is often out of the experimenter's control, in the context of advertising e□ectiveness studies, some platforms report e□ectiveness at the user level instead of the cookie level and advertisers should keep this in mind when comparing results between platforms. Given similar e□ect sizes between platforms, more budgetmustbe allocatedtocookie-level adsystems to find statistically significant sales lifts.

Some extensions may provide further insight into the magnitude of this problem. Interesting directions for future work include relaxing the assumption of independence of treatment and spending per cookie conditional on the number of treated cookies; allowing for the number of cookies a user generates to be a□ected by previous cookies' treatment status; richer treatment e□ect frameworks, which nest both the case we consider and the case in which there are *no* spillovers between cookies; and allowing heterogeneity in treatment effects across the number of total cookies.

6. ACKNOWLEDGMENTS

Many thanks to the Atlas Marketing Team, Eytan Bakshy, Tom Blake, Brad Larsen, Randall Lewis, Garrett Johnson, Dimitriy Masterov, Kane Sweeney and Steve Tadelis for helpful comments anduggestions on this paper.

References

[1] A. Abadie, J. Angrist, and. Imbens. Instrumental variables estimates of the e□ect of subsidizedtraining on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002.

[2] J. D. Angrist, G. W. Imbens, and B. Rubin. Identification of causal e□ects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[3] E. Bakshy, D. Eckles, R. Yan, and Rosenn. Social influence inocial advertising: evidence from field experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 146–161, 2012.

[4] A. Balke and J. Pearl. Bounds on treatment e□ects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[5] T. Blake and D. Coey. Why marketplace experimentation is harder than it seems: the role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582. ACM, 2014.

[6] T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search e□ectiveness: A large scale field experiment. *Econometrica*, 2015.

[7] P. Chatterjee, D. L. Ho□man, and T. P. Novak. Modeling the clickstream: Implications for web-based advertising e□orts. *Marketing Science*, 22(4):520–541, 2003.

[8] A. Dasgupta, M. Gurevich, L. Zhang, B. Tseng, and A. O. Thomas. Overcoming browser cookie churn with clustering. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 83–92. ACM, 2012.

[9] X. Dreze and F. Zufryden. Is internet advertising ready for prime time? *Journal of Advertising Research*, 38:7–18, 1998.

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

[11] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 81–87. ACM, 2010.

[12] J. Hadar and W. R. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, pages 25–34, 1969.

[13] J. J. Heckman, S. Urzua, and E. Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

[14] J. J. Heckman and E. J. Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007.

[15] P. R. Hoban and R. E. Bucklin. E□ects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3):387–393, 2015.

[16] G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment e□ects. *Econometrica*, 62(2):467–475, 1994.

[17] G. A. Johnson, R. A. Lewis, and E. I. Nubbemeyer. Ghost ads: Improving the economics of measuring ad e□ectiveness. 2015.

[18] G. A. Johnson, R. A. Lewis, and D. H. Reiley. Location, location, location: Repetition and proximity increase advertising e□ectiveness. 2015.

[19] A. Klenke, L. Mattner, et al. Stochastic ordering of classical discrete distributions. *Advances in Applied probability*, 42(2):392–410, 2010.

[20] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM, 2012.

[21] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967. ACM, 2007.

[22] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[23] A. Lambrecht and C. Tucker. When does retargeting work? information specificity in online advertising. 2013.

[24] R. A. Lewis and J. M. Rao. The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics (forthcoming)*, 2015.

[25] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the e□ects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157–166. ACM, 2011.

[26] R. A. Lewis and D. H. Reiley. Does retail advertising work? measuring the e□ects of advertising on sales via a controlled experiment on yahoo! 2010.

[27] R. A. Lewis, D. H. Reiley, and T. A. Schreiner. Can online display advertising attract new customers? 2010.

[28] P. Manchanda, J.-P. Dubé, K. Y. Goh, and P. K. Chintagunta. The e□ect of banner advertising on internet purchasing. *Journal of Marketing Research*, 43(1):98–108, 2006.

[29] M. Rothschild and J. E. Stiglitz. Increasing risk: I. a definition. *Journal of Economic theory*, 2(3):225–243, 1970.

[30] N. Sahni. Advertising spillovers: Field-experiment evidence and implications for returns from advertising. 2013.