Incrementality Testing in Programmatic Advertising: Enhanced Precision with Double-Blind Designs

Joel Barajas
Yahoo Research, Verizon Media
Sunnyvale, CA, USA
joel.barajas@verizonmedia.com

Narayan Bhamidipati
Yahoo Research, Verizon Media
Sunnyvale, CA, USA
narayanb@verizonmedia.com

ABSTRACT

Measuring the incremental value of advertising (incrementality) is critical for financial planning and budget allocation by advertisers. Running randomized controlled experiments is the gold standard in marketing incrementality measurement. Current literature and industry practices to run incrementality experiments focus on running placebo, intention-to-treat (ITT), or ghost bidding based experiments. A fundamental challenge with these is that the serving engine as treatment administrator is not blind to the user treatment assignment. Similarly, ITT and ghost bidding solutions provide greatly decreased precision since many experiment users never see ads. We present a novel randomized design solution for incrementality testing based on ghost bidding with improved measurement precision. Our design provides faster and cheaper results including double-blind, to the users and to the serving engine, post-auction experiment execution without ad targeting bias. We also identify ghost impressions in open ad exchanges by matching the bidding values or ads sent to external auctions with held-out bid values. This design leads to larger precision than ITT or current ghost bidding solutions. Our proposed design has been fully deployed in a real production system within a commercial programmatic ad network combined with a Demand Side Platform (DSP) that places ad bids in third-party ad exchanges. We have found reductions of up to 85% of the advertiser budget to reach statistical significance with typical ghost bids conversion and winner rates. Moreover, the highest statistical power at 50% control size design of this current practice is reached at 8% of our proposed design. By deploying this design, for an advertiser in the insurance industry, to measure the incrementality of display and native programmatic advertising, we have found conclusive evidence that the last-touch attribution framework (current industry standard) undervalues these channels by 87% when compared to the incremental conversions derived from the experiment.

CCS CONCEPTS

• **Applied computing** → **Marketing**; **Economics**; • **Information systems** → **Computational advertising**; **Display advertising**; **Online auctions**; • **General and reference** → **Experimentation**; • **Mathematics of computing** → **Probability and statistics**.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450106

KEYWORDS

Marketing Incrementality, Ad Effectiveness, Controlled Randomized Experiments, A/B Testing, Computational Advertising, Programmatic Advertising, Causal Inference

ACM Reference Format:
Joel Barajas and Narayan Bhamidipati. 2021. Incrementality Testing in Programmatic Advertising: Enhanced Precision with Double-Blind Designs. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3442381.3450106

1 INTRODUCTION

Measuring the incremental value of advertising is a critical estimation for advertisers for financial planning and optimal budget allocation [15]. Advertisers often run a portfolio of marketing media channels, often characterized by the media format, *e.g.* native, display, social, sponsored search, video, among others. In these heterogeneous advertising ecosystems, relying on standard last-touch attribution poses significant issues. That is because there is a bias towards channels closer to conversion in the funnel (demand capture channels), such as sponsored search, versus demand generation channels, *e.g.* programmatic native or display [19]. Multi-touch attribution modeling attempts to address this problem [22], including theoretical modeling approaches [5, 23]. However, these methods generally assign credit to every exposed and converting user while ignoring the counterfactual response of the user without ad exposures.

Finding the incremental value of marketing budgets (*incrementality*) translates into measuring the causal effect of the marketing budget on a target metric. Typical metrics include: sales, acquisition, brand awareness, among others. In this context, identifying the causal value of ad exposures requires finding the counterfactual user response of those exposed to ads when no ad would have been displayed, aggregated for all exposed users. This problem has been addressed from logged observational data post-campaign [8, 12, 25]. Nonetheless, Gordon *et al.* concluded that these methods, even in the presence of rich user-level data, do not provide a reliable estimate of the advertising effectiveness [11].

The most widely accepted approaches by the research community and the ad tech industry are to measure incrementality using randomized experiments (A/B testing) [3, 7, 10, 13, 18]. At a fundamental level, running an experiment for incrementality measurement sets aside without ads a randomly selected hold-out group of users (control group). This group provides a counterfactual view of the user response without the ad. However, due to the complexity of modern advertising serving systems, the experiment execution is not trivial. That is mainly because identifying counterfactual

user impressions in the hold-out group is extremely challenging. Without ads in the control group, the key challenge is to precisely identify the counterfactual ad exposed users [3, 13].

Current research literature and industry practices to run incrementality experiments focus on running: placebo, intention-to-treat (ITT), or *ghost* would-be bidding based experiments. Placebo based experiments require running parallel campaigns in the control group showing unrelated ads, *e.g.* charity ads [18]. ITT based solutions do not touch the control group resulting in the most unbiased campaign counterfactual [3]. Ghost bidding based solutions rely on the ability to log would-be (ghost) ad opportunity events at the last controllable step by the ad serving engine, typically submitted bids to ad exchanges [7, 13]. The fundamental challenges with these solutions in the literature are: 1) the serving engine as treatment administrator is not blind to the user treatment assignment, 2) ITT and ghost bidding solutions results in large variability in the effect estimations leading to a decreased precision.

1.1 Our Contribution

We present a novel experimental design solution for incrementality studies based on ghost ads with improved measurement precision. Our design provides faster and cheaper results including:

- Truly double blind (to the users, and to the targeting system) post-auction experiment execution without typical ad targeting bias.
- Identifying ghost impressions in open ad exchanges by matching the bidding values with held-out bid values, leading to larger precision than ITT studies.
- Simple transparent statistics from observed data without major modeling or tuning.

For ad call traffic within programmatic ad networks, our design controls the serving marketplace and executes the experiment split post-auction, and before pricing and rendering the ad. With post-auction randomization, the ad as the experiment treatment is delivered blindly without any targeting or auction bias. For real time bidding (RTB) programmatic traffic, we execute the experiment split at bidding time by matching bidding prices of the alternative ad sent to external exchanges. Both solutions combined provide the experiment precision at the user impression level, not at the bid pre-auction level as performed by current ghost bidding solutions.

Our solution reduces the cost of experimentation to the bid differences in RTB traffic. It further eliminates the cost within ad networks traffic completely. As a double blind design, it is suitable to test *any* single or mixed ad targeting strategy, such as re-marketing or prospecting marketing, without any constraint. Also, the proposed design greatly increases precision and statistical power when compared to ITT practices in industry [7, 10]. As a result, the causal effect estimation is derived from simple transparent statistics without major modeling or tuning.

This randomized design has been fully deployed in production within a commercial programmatic ad network, and in a commercial Demand Side Platform (DSP) that places ad bids in third party ad exchanges. To our knowledge, our approach is the first design being deployed in a production system that logs counterfactual ghost impressions within ad network and RTB traffic.

Leveraging this design to measure the incrementality of display and native programmatic advertising for an advertiser in the insurance industry, we provide conclusive evidence that the last-touch attribution framework (current industry standard) undervalues these channels when compared to the incremental conversions value derived from the experiment.

2 LITERATURE REVIEW

Measuring advertising channel incrementality has been previously addressed as an online conversion attribution problem. Shao and Li approached multi-touch attribution as a feature importance machine learning problem regardless of the nature of the touch point [22]. By addressing the fundamental difference of touch points between user-initiated and firm-generated Li and Kannan developed an econometric based model to attribution [19]. More fundamental attribution models have been developed based on incentives to publishers and advertisers [5], and Markovian user stage modeling in the conversion funnel [23]. Fundamentally, these approaches generally assign credit to every exposed and converting user without considering the user response without any ad [3].

Identifying the causal value of ad exposures requires finding the counterfactual user response of exposed users when no ad is displayed to them. One stream of research to solve this problem is to estimate this counterfactual conversion rate response from logged observational data post-campaign. In online advertising evaluation, these approaches range from propensity-scores based approaches [8], to more sophisticated frameworks to approximate user response heterogeneity [12, 25]. However, typical advertising effects on conversion rates come with small lift effects and sparse conversions. Also, modern ad serving systems oftentimes include sophisticated machine-learning based targeting and bidding strategies in marketplaces. As a result, observational methods become too sensitive to the control features and errors [3, 18], even in the presence of rich user-level data [11].

Geo-testing based approaches are deployed in cases where user-level experiment execution is not feasible to advertisers. These approaches rely on geo-targeting to define treat/control geo-units, *e.g.* US DMAs[1], and estimate effects based on aggregate time series models [4, 6, 14]. The fundamental constraint with these approaches is the large variability and low number of geo-units. To address this problem, Barajas *et al.* have proposed to improved both, the experimental design and the causal estimation [4] with synthetic-control based methodologies [1]. However, these methods are still sensitive to holidays and geo-specific events, such as major events. Consequently, these methods pose significant constraints to specific time-of-year testing periods and require a long experiment duration.

We review incrementatlity testing practices and literature in the next section.

3 INCREMENTALITY TESTING PRACTICES

We approach the ad incrementality measurement using randomized controlled experiments (A/B testing), which is summarized as:

Goal: Find the aggregate effect of advertising spend
Randomized Units: Users

[1] Nielsen DMA Regions. https://www.nielsen.com/intl-campaigns/us/dma-maps.html

Figure 1: User groups observed for current incrementality measurement practices. From left to right, placebo campaign approaches [18], ITT approaches including ghost bidding [3, 13], our approach based on ghost impressions.

Intervention: Advertising spend leading to the delivery of ad impressions

Control: No ad impressions

We refer to *treat* group as the group of users who are exposed to the intervention, and *control* as the group of users who are not exposed to the intervention.

Without impressions in the control group, the key challenge is to identify the counterfactual ad exposed users. Research literature to address this problem using experiments can be divided into:

- Placebo based approaches
- Intention-to-treat based approaches
- Ghost-events based approaches

In placebo based approaches, a parallel campaign is launched with the same settings but displaying unrelated ads, such as public service announcements or PSAs, to a randomly assigned control group [18]. However, PSA based testing suffers from potentially different audience selection bias [3, 13], and requires a significant cost when delivered for the entire advertiser channel [18]. Figure 1 illustrates this bias. Fundamentally, placebo campaigns do not score and bid on ad calls the same way as the actual campaigns do, violating the necessary user selection blindness of the serving engine during treatment intervention. This constraint is particularly problematic for ad campaigns that rely on user feedback, *e.g.* re-marketing, where targeting models are updated based on this feedback. Since the placebo campaigns do not receive the same feedback their models are not updated the same way, introducing a selection bias.

Intention-to-treat (ITT) based approaches rely on the selection of users who could be potentially exposed in both treatment and control groups, *i.e.* reachable users. In this design, the control group of users is not influenced in any way reflecting the accurate counterfactual of the complete absence of the ad campaign as depicted by Figure 1. Often the challenge in these designs is to determine the user experiment qualifying event (or exposure logging event in product experimentation platforms [16]). One approach is to make the user visit to the set of publisher pages where the campaign ads are displayed in the treat group [3], or at user segment level to avoid accounting for the entire user base within ad networks [10]. The fundamental problem with this solution is the experiment precision with large variability leading to great challenges achieving statistically significant estimations of the effects [17] – typical exposure rates are less than 30% in 1 month of testing [3].

Ghost based approaches rely on the ability to run the ad delivery process in the control group up to the last controllable point in the ad serving process. At this point, control users are held out and a *ghost* event is logged. One of the closest applications of this practice is Ghost Bidding [13], where the experiment execution event is the bid response from DSPs to third-party exchanges. At that point, when the focal advertiser's ad bid is above to be submitted to the ad exchange (internal DSP auction winner), the bid is held out and replaced by the next ranked ad bid. Then, the ghost bid event is logged. Since not all the submitted bids win, this Ghost bidding design becomes a variant of ITT designs leading to further modeling and diminished precision [7]. Johnson *et al.* suggest simulating auctions (ITT analysis), assuming that external auctions can not be won and repeated for winner bids in the control group [13].

Compared to placebo based testing, our proposed design is double-blind to the user and the targeting engine eliminating the potential targeting bias. Since our randomization checking point happens *after* ad scoring and ranking, targeting models are blind to the randomized experiment. Compared to typical ITT based approaches with pre-randomized users, we rely on ghost impression logging. We contrast ghost impressions in our approach, which are bid-and-won counterfactual impressions, with ghost bids in current industry practices, which are limited to held-out bid calls only. We solve the ghost impression logging problem within ad network traffic by running the auction for more ads than requested, and randomizing users post-auction. When we interact with third-party ad exchanges as a DSP, we match the bid value of the ghost bid to the next-ranked ad in our internal auction that is sent to the exchange. If that bid is won, we observe the ghost impression. Combine, these conditions provide a double-blind experimental design at the impression-level precision of placebo targeting.

4 METHODOLOGY

In this section, we first review the typical ad serving flows within Ad Networks and DSPs. We then define the randomized experiment design, and describe the causal estimation framework. Finally, we review standard attribution metrics for comparison purposes with incrementality metrics.

4.1 Ad Network and DSP Typical Serving Flows

Before identifying the right randomized design that guarantees being blind to the ad serving engine, we review typical ad serving components for ad network traffic and RTB traffic from the DSP perspective. Figure 2 shows the high-level serving flows.

For a given ad request, in both ad network traffic and RTB traffic, there is generally a targeting and scoring component. Here is where the ad network or the DSP enforces target segments and deploys predictive modeling for a given advertiser goal. We note that this component is the source of most of the selection bias when placebo ad campaigns are deployed. For ad networks traffic, there is an auction process within its marketplace that ultimately selects the winner ad based on a set of optimized submitted ad bids.

Serving ads for RTB traffic implies interacting with third-party ad exchanges. In contrast with traffic within ad networks, DSPs often have an internal auction, an additional step, in the ad serving

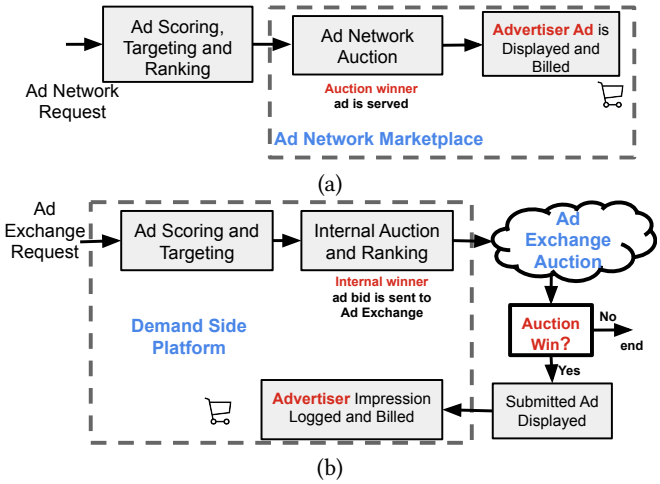

Figure 2: Typical ad serving workflow for: (a) Ad Network traffic, (b) RTB traffic for DSP ad serving engines.

process before submitting bid responses. Then, the exchange auction mechanism determines the winner bid, responds to the DSP if the submitted ad is displayed, and bills the DSP for the impression.

Within these flows, the point where we check for treat or control user assignments determines what is blind to the experiment. Since users in both groups will be subject to the same flow regardless of their assignments before that point, any process behind it is blind to the experiment.

4.2 Randomized Controlled Design

4.2.1 User Treatment Assignment. Given the most reliable user identifier available at the ad call time (*e.g.* cookie, device id, or vendor specific identity id), we randomly assign users into treat and control groups. Treat users are exposed to regularly-delivered ads based on any targeting strategic goal. Control users are prevented from being exposed to any treat ad. We note that for optimal budget allocation at the channel level, the ad exposures treatment corresponds to all ads delivered through the channel over a period of time. As a result, the user treatment assignment needs to:

- Randomize effectively new users as they join the experiment
- Stick to a user consistently over the duration of the test

To achieve this goal, we use a hashing function evaluated at serving time, which uses the user id and the experiment id as combined key[2]. This design guarantees both conditions above as never-seen users will automatically be assigned to a group and will consistently fall in this group for future ad calls.

We contrast this design setup to other practices where users are randomized first and the experiment becomes available only to those focused user cohorts. This practice poses a potential selection bias as previously randomized users are often more active than new users (hence their identification before becoming eligible to see ads). We argue this user selection is not representative of the regular spending with significant effects on the external validity of the experiment results.

4.2.2 Traffic within Ad Networks. As we discussed in Section 4.1, the bifurcation point where the serving flow identifies the user treatment assignments is critical to set the blindness of the serving engine as treatment administrator. For traffic within ad networks, where the marketplace is within control of the network, we set the user hashing post-auction as depicted by Figure 3(a).

In this flow, we execute the randomized design for each ad call where the advertiser's ad is ranked by holding ads for control users post-auction and before pricing. We request additional ads to the auction (a system parameter), and serve the next ranked ad if the user is assigned to the control group. For treat users we serve the highest ranked ad as usual. If the served ad is rendered we log the ghost impression (control) or impression (treat). Since the users are randomized post-auction, this design is blind to the users and to the serving engine as treatment administrator (double-blind). We note that the likelihood of multiple control group users (from any concurrently running test), at a given ad call decreases exponentially to the control group size. For instance, if four advertisers are running their experiments concurrently with each having a control group of 10%, the probability of encountering 4 ghosts at serving time is 0.01%.

This design guarantees the correct logging of control ghost impressions, which are directly comparable to treat regular impressions. As a result, we achieve the goal of experiment precision to the level of users exposed to the ad impression. Compared to placebo based testing, we achieve the same precision without selection bias risk with a double blind design and without the cost of PSA ads. Compared to ghost bidding and ITT studies, we achieve the same unbiased results but with greatly increased precision by discarding users with bids that are not won.

4.2.3 Third-Party Exchanges: RTB Traffic. For RTB traffic, identifying ghost impressions is more complex given that DSPs interact with third-party ad exchanges. Thus, the marketplace is not within the control of the DSP running the incrementality test. This constraint is discussed by Johnson *et al.* suggesting simulating auctions (ITT analysis) assuming it is not possible to go beyond that point in the DSP serving flow [13].

We execute the experiment at the time of sending the bid response to the ad exchange as depicted by Figure 3(b). In this process, for ad calls coming from programmatic ad exchanges that the advertiser's ad is the DSP internal auction winner, we hold out control users (randomly hashed) before submitting bids to external auctions. We then match the bid value of the held-out ad to the next internal winner sent to the exchange. If the bid wins the auction and the ad is displayed we log the ghost impression (control) or impression (treat). As the average probability of winning the exchange auction is largely a function of the bid price, we are able to log the ghost impressions. This randomized design provides an unbiased user selection based on the following assumption:

Assumption 1. *Auctions are approximately content-blind. The probability of winning at external auctions is the same, in the average, for equal bid prices.*

[2]With offline experiments, we validated the hashing function for different types of ids, which has been addressed by prior work in product experimentation platforms [16].

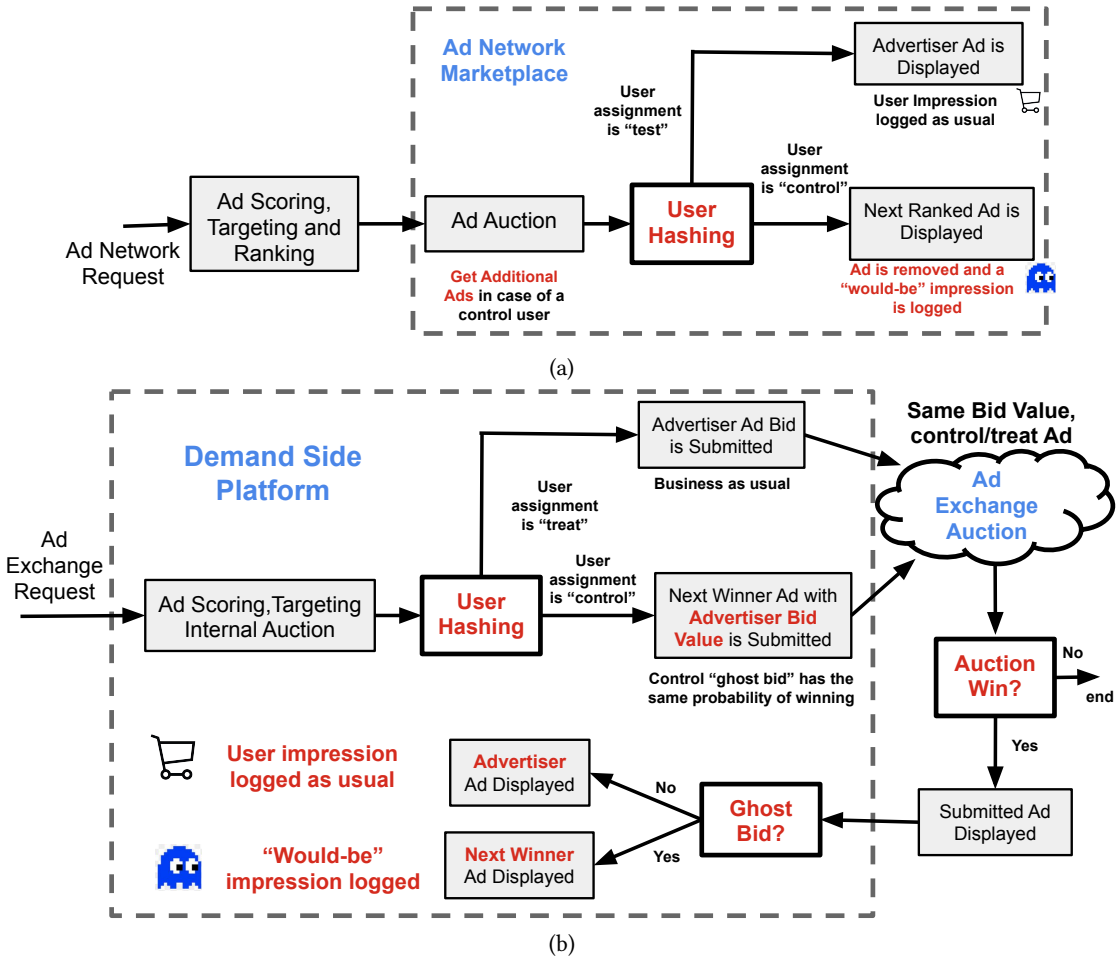

Figure 3: Randomized experiment execution process for (a) Ad Network traffic, (b) RTB traffic for DSP ad serving engines

We note that Assumption 1 is based on the auction competition, regardless of the bidding DSP policies, e.g. predictive click-through rate or predictive conversion rate. By sending the second-ranked ad to the exchange with the same bid value, the proposed design transfers the ad policy value originally set on the held-out ad.

By logging ghost impressions, we eliminate users who have no effect but whom would need to be included in the ITT analysis of standard ghost bidding as depicted by Figure 1. We note that typical winning impression rates are often less than 10%. Also, user exposure rates have been reported to be less than 30% in 1 month of testing [3], leading to 70% of users who never see any ad because none of the bid auctions to reach them was won. All these users need to be included in the ITT analysis increasing its variability and noise.

These designs provide parity to users exposed to ads in both ad network and third party ad exchanges traffic. Consequently, control users are directly comparable to treat users for both ad network and RTB traffic. We maintain the same properties described above for ad network traffic, that is placebo based precision with double-blind unbiased designs, aligned to the RTB traffic. As a result, based on a common experiment identifier and user id, we align the experiment execution in both traffic sources providing an aggregate view of the channel media.

We note that the randomization point is similar to ghost bidding approaches [7, 13]. However, by matching the bid of the next internal winner ad with the bid of the held-out user the precision is dramatically improved. In the proposed randomized design, we use the term *matching* as using the winning bid for the next-ranked ad in the control group. There is no relation to the typical usage of matching in causal inference, e.g. propensity score matching. The source of additional statistical precision comes from observing the control group of users, those who are logged with ghost impressions, to those who actually see ads without any inference (see ITT vs Ghost Ads in Figure 1). In Section 5.1, we quantify the precision gains with a number of randomized experiment scenarios.

4.2.4 A Note on the Cost. Bidding the runner-up of the internal auction with the bid of the ghost ad is bound to increase the cost paid in the external auction. Naturally, questions would arise regarding who bears this additional cost. Here are some plausible options:

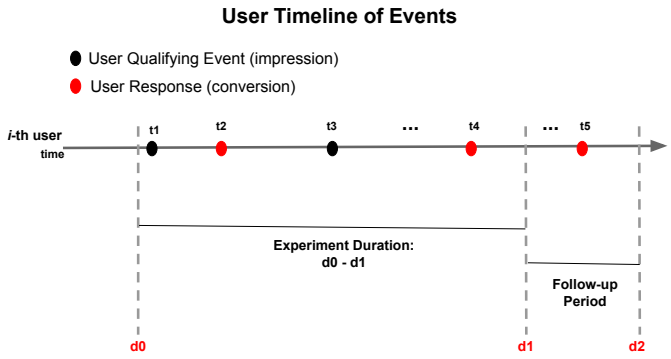

Figure 4: User timeline of events to determine experiment cohorts and responses. Qualifying event to join the experiment cohort: first impression/ghost impression (t_1). User response: conversions after that event ($> t_1$).

- Advertiser running the experiment bears the cost. This is similar to them paying for the cost of placebo PSA ads, although this would be philosophically different. In particular, the advertiser may feel that this cost is subsidizing the campaigns of other advertisers, who get access to more expensive audiences without incurring the additional cost. Note that for sufficiently large ad networks with diverse audiences and advertisers, the cost would likely be spread out over several advertisers, so no particular advertiser would appear to be a huge benefactor. In our experience, advertisers are willing to pay for a reliable solution given that their alternatives on other platforms are neither free nor accurate.
- The increased cost, or at least part of it, could be borne by the runner-up advertiser from the internal auction. Most optimization systems would need to perform bid exploration to better determine the true bid landscape and where exactly a campaign would see optimal performance. Typical exploration scenarios involve bidding higher than the original estimate the system produced, so at least part of the increased cost may be recovered via the exploration costs. Note that, once again, no single advertiser is expected to be the runner up each time.
- Finally, the ad network may absorb the cost using one of several justifications. For example, this could be considered the cost of doing business. Offering such precise measurement solutions would attract more advertisers and increased budgets resulting in a net gain. Advertisers could also increase their bids once they realize that this ad network is being undervalued based on the last-touch attribution model. Another potential option is for the ad network to strike an all-you-can-eat private deal with the advertiser for an upfront payment as a cost of running the test regardless of how many ghost impressions or cost differential the experiment would eventually have. Of course, the ad network could use a mix of all of the above to keep costs below a threshold for all the players involved.

4.3 Causal Estimation

As in any randomized experiment and adhering to concepts from product experimentation platforms [16], we identify the user responses after they join the experiment based on their eligibility event. This event is the first advertiser impression/ghost impression, and the user responses are all conversions (or any other relevant user response) after that event. Figure 4 depicts this process.

In the jargon of online advertising, this framework is equivalent to *any touch* user conversion join in both treatment groups. From an experimentation point of view, user responses could be any metric and vary based on the advertiser specific needs. Note that the goal of the experiment is to find the user counterfactual conversion response of users who are exposed to ads. This response is equivalent to the organic probability of user conversions without the ad treatment exposures. Thus, standard constraints with last-touch attribution are not applicable.

Given this data collection, estimations are straightforward. We illustrate the effect on converters as a metric (users with one or more conversions) for simplicity, but the same calculations apply to other metrics including: conversions, sales, acquisition sign-ups, advertiser on-boarding funnel stages, etc).

Within the Potential Outcomes causal framework [21], we define the following indicator random variables for each user i: Z_i for control/treat group user random assignments $\{0, 1\}$, Y_i for non-converting/converting users $\{0, 1\}$. Let $n_{Y=y,Z=z}$ be the count of unique users who met the conditions $Y_i = y, Z_i = z$. Thus, we define the average treatment effect on exposed users ATE_{ads}, lift $lift_{ads}$, incremental converters generated by the advertiser spend $ATRB_{ads}$, and the cost per incremental converter $CPIA$ as follows:

$$
\begin{aligned}
ATE_{ads} &= E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0], \\
lift_{ads} &= ATE_{ads}/E[Y_i|Z_i = 0], \\
ATRB_{ads} &= ATE_{ads} \times n_{Z=1}, \qquad CPIA = Spend_{ads}/ATRB_{ads},
\end{aligned}
\tag{1}
$$

where $Spend_{ads}$ is the total spend to deliver those ads.

Given the scale of a typical test, i.e. number of experiment units, statistical confidence intervals are estimated using the Central Limit Theorem from aggregate experiment unit counts, response counts, and sum of squared responses. Similar to a two-sample t-test and assuming Bernoulli likelihood distribution of Y_i, we estimate:

$$
\begin{aligned}
ATE_{ads} &\sim N\left(\bar{Y}_{Z=1} - \bar{Y}_{Z=0}, \frac{S^2(Y_{Z=1})}{n_{Z=1}} + \frac{S^2(Y_{Z=0})}{n_{Z=0}}\right), \\
\bar{Y}_Z &= \frac{n_{Y=1,Z}}{n_{Y=0,Z}}, \qquad S^2(Y_Z) = \bar{Y}_Z \times (1 - \bar{Y}_Z).
\end{aligned}
\tag{2}
$$

Extending this estimation to the case of multiple conversions per user, $Y_i \in \{0, 1, 2, ...\}$ or a continuous response (e.g. revenue), is trivial by estimating the sample mean, $\bar{Y}$, and the sample variance $S^2(Y)$ for both user groups.

We note that the units used for the causal effect estimation must be aligned with the experiment units (per Potential Outcomes condition), which in our design are users. Estimations based on other units, e.g. impressions, would potentially be biased since the experiment assignment mechanism is *ignorable* to user features without further stratification or blocking factors [21].

Table 1: Precision gain between proposed design and the literature ITT design for ghost bidding with ad exchanges for different detectable lifts. Control group size: 10%. Reachable users, those that we bid: 100M. Confidence level: 95%.

Minimum Detectable Lift	Converter Rate				Ad Exposed Users Needed		Budget Gain
	Control Group		Treat Group				
	Ad Targeted	No Ad Targeted	Ad Targeted	No Ad Targeted	ITT Design	Proposed Design	
15%	0.135%	0.05%	0.155%	0.05%	8M	1.28M	**84%**
10%	0.135%	0.05%	0.149%	0.05%	12M	2.80M	**76%**
5%	0.135%	0.05%	0.142%	0.05%	27M	11.40M	**58%**
3%	0.135%	0.05%	0.139%	0.05%	49M	31.67M	**35%**

4.4 Click-to-Conversion Last-Touch Attribution

Last-touch attribution has been the standard value attribution in online advertiser for more than a decade, despite numerous studies highlighting its issues [5, 22, 23]. Scientific evidence has shown that last-touch attribution based on viewed impressions over-estimates the value of display advertising [18]. As a result, industry has adopted a click-to-conversion attribution model (C2C) in an attempt to diminish this over-estimation. In this attribution framework, the conversion value is attributed to the last click in the user path (i.e. post-click conversions). Few studies have been conducted to assess the value of overall online clicks [24] or ad clicks [2, 9].

As a form of comparison with the incremental conversions and their cost, we define:

- $ATRB_{C2C}$: Last-touch post-click conversions based on C2C business model
- CPA_{C2C}: Cost per attributed conversion based on C2C business model

As a result:

$$CPA_{C2C} = Spend_{ads}/ATRB_{C2C}.$$

Comparisons between CPA_{C2C} and $CPIA$ provides evidence if the C2C business model over-estimates the channel attribution (a well-accepted belief for most forms of online advertising) or not.

5 RESULTS

5.1 Value of Precision Enhancement

Identifying counterfactual impressions in the control group allows us to perform a simple mean difference effect estimation without noise from users with no ad exposures. We quantify these benefits in terms of the minimum exposed target users necessary to measure a statistically significant effect when compared to the ghost bidding pre-auction ITT design [7, 13]. We use the Local Average Treatment Effect (LATE) estimation deployed in incrementality studies before [3] to simulate effect scenarios in a statistical power analysis. For this analysis, we run test simulations (offline A/A tests) for a given set of parameters and user populations. Thus, assuming known parameters and a minimum detectable lift, we compare the ghost bidding pre-auction ITT design to our proposed design, and estimate user populations needed for a statistically significant estimation.

Table 1 shows the power analysis results for different minimum detectable lift scenarios with typical converter rates of 0.05% and 0.135% for no-ad targeted and ad targeted groups, respectively. Assuming 100M reachable users with 95% confidence, we observed budget gains of up to 84% for a 15% minimum detectable lift when compared to the ghost bidding ITT design. Even for a minimum detectable lift of 3%, the budget gain represents a 35%. Note that the savings arise from the fact that ITT analysis can only compare the converter rates across the entire control and treat groups, so the converter rates being compared are diluted, whereas our design compares the converter rates for the undiluted Ad Targeted populations within control/treat groups.

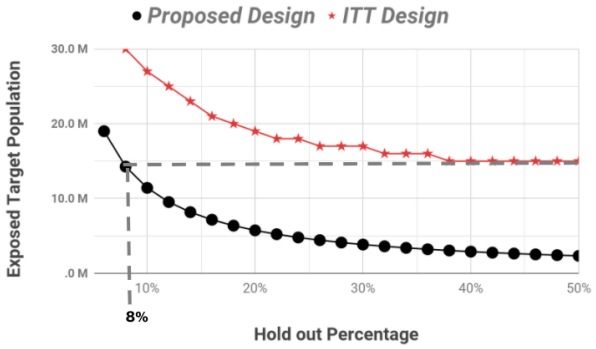

Figure 5: Exposed target population as a function of the control group size for the proposed design (circle marks) and the ITT ghost bidding design (star marks). Control converter rate: 0.135%. Minimum detectable lift: 5%. Reachable users: 100M. Confidence level: 95%.

Considering the parameters of Table 1 for a minimum detectable lift of 5%, Figure 5 shows the exposed target population size as a function of different control group percentages. The population target size converges in both methods in a similar shape but with a clear offset gain in experiment precision. We note that the highest statistical power detection scenario for the **ghost bidding ITT design at 50% control size is reached at 8% control group of the proposed design**. This scenario represents a target population of 15M users exposed to ads. As depicted, the proposed design converges to 2.3M users at 50% control group representing a budget gain of 85%.

5.2 RTB and Native Ad Network Test Results

We ran an actual experiment in a commercial production system deploying our design for traffic from a Native Ad Network and from

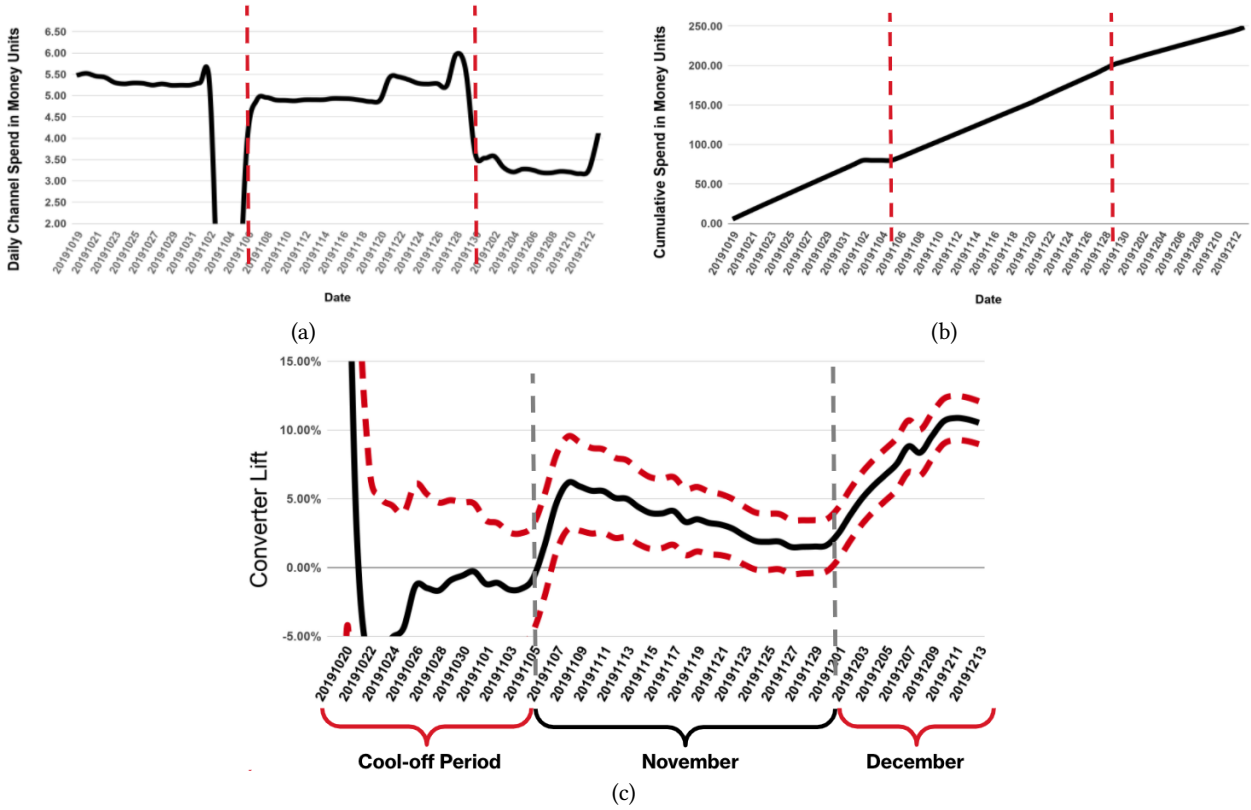

Figure 6: Incrementality test result progression for an insurance industry advertiser in the US with combined traffic of a commercial Native Ad Network with a DSP placing ads in RTB traffic. (a) Daily channel spend in money units. (b) Cumulative spend over the duration of the experiment. (c) Cumulative converter lift progression. 95% confidence intervals are displayed. Conversion definition: insurance quote. Duration: 8 weeks, 10/18/2019 - 12/12/2019. Holdout size: 10%.

third-party RTB exchanges, for an insurance industry advertiser. The test use case falls in the optimal budget allocation category [15], where the advertiser's goal is to compare incremental performance for different media channels. The primary metric of the test is number of online insurance quotes. The advertiser has been regularly running ad campaigns and the incrementality test execution would start holding out control users.

Figure 6 shows the cumulative converter lift evolution with 95% confidence intervals, the daily spend pattern and the cumulative spend progression[3]. We observe a *cool-off* period coming from regular ad spending, which reflects any carryover effect on control users. During this period of 2.5 weeks, the effects vary and move greatly as more users joins the experiment.

As discussed in 4.2, the design is blind to any targeting or strategic goal in the campaign management. This property allows us to estimate the channel incrementality even when the advertiser executes different strategies. In the result progression, we observe that the advertiser re-adjusted strategies and campaigns every month. Also, we did not have any constraint with running the test during holidays (a typical constraint in geo-testing [4]). These results illustrates the benefits of the double blind randomized design.

Overall, during 8 weeks of testing, we detected a user converter lift of 10.5% +/- 1.56% as listed in Table 2. We also found conclusive evidence that the industry standard last-touch attribution undervalued the contributions of display and native programmatic advertising as a combined channel by 87% compared to the incremental conversions value derived from the experiment. We observed that only **13% of incremental conversions were attributed to the channel** based on C2C post-click conversion attribution business model. This findings pose a stark contrast with the common belief that last-touch attribution over-values the effects of online advertising, and are consistent with previous literature [2].

6 CONCLUSION AND MANAGERIAL IMPLICATIONS

We have introduced two randomized designs to improve the incrementality measurement precision of programmatic advertising within ad networks and in third-party ad exchanges. Our designs achieve the same precision as placebo based testing without any

[3]For business privacy reasons, we report the spend in *money units*, which is a scaled number of the actual channel spend.

Table 2: Aggregate lift and results, and comparison with last-touch C2C business model attribution for the test progression of Figure 6. 95% confidence intervals (*Low,High*) are displayed. Conversion definition: insurance quote. Duration: 8 weeks, 10/18/2019 - 12/12/2019. Holdout size: 10%

Converter Rate Lift			Conversions per User Lift			$CPA_{C2C}/CPIA$	$ATRB_{C2C}/ATRB_{ads}$
Low	Average	High	Low	Average	High	Rate	(%)
8.98%	10.54%	12.10%	6.82%	9.51%	12.20%	**7.58**	**13.20%**

selection bias. With double blind designs, we eliminate any constraint to the targeting or strategic optimization goal by online campaigns. The improved precision has demonstrated significant gains, which translate into faster testing and consequently more effective decision making.

We have provided evidence that post-click conversion attributed value (current industry standard) greatly underestimates the value of marketing spend. Although attribution modeling is focused on making heterogeneous media adverting types comparable, our findings suggest that industry attribution practices favor channels lower in the conversion funnel (*e.g.* paid search). Overall, these findings raise questions about the success of attribution business models as a one-fits-all solution.

Incrementality, as a targeting optimization goal, suggests showing ads to *persuadable* users. These are users who convert in the treat group but would not convert in the control group. However, executing this goal comes with a significant cost in precision. Since we can not observe who is a persuadable user, we can not deploy predictive machine learning algorithms out of the box. As a result, more modeling errors are introduced to the already-sparse user conversions. Being able to test any advertising strategy (per double blind designs) with improved impression-level precision opens the door to incrementality driven targeting optimizations.

We randomize users, which translates to user identifiers. These identifiers range from cookies or device ids to cross-device identifiers (*e.g.* logged-in user identifiers or email addresses). We emphasize the importance of these identifiers to effectively separate treatment experiences among vendor-specific identity graphs. Recent work by Lin and Misra addresses the challenges of identity fragmentation and potential biases within observational studies [20]. For the case of our randomized designs, fragmented identities translate into a decrease in experiment precision due to unobserved spillovers. With increasing challenges to unify user identities from also increasing heterogeneous sources, this problem is emerging as an open research topic.

Along similar lines, with increasing restrictions on privacy, the number of trackable users is expected to significantly decrease over the next few years. Consequently, parsimonious designs such as the one presented here will have a major advantage as they require fewer users and impressions to arrive at statistically significant conclusions regarding the causal effects of advertising.

REFERENCES

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association* 105, 490 (2010), 493–505.

[2] Joel Barajas, Ram Akella, Aaron Flores, and Marius Holtan. 2015. *Estimating Ad Impact on Clicker Conversions for Causal Attribution: A Potential Outcomes Approach.* 640–648. https://doi.org/10.1137/1.9781611974010.72 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611974010.72

[3] Joel Barajas, Ram Akella, Marius Holtan, and Aaron Flores. 2016. Experimental designs and estimation for online display advertising attribution in marketplaces. *Marketing Science* 35, 3 (2016), 465–483.

[4] Joel Barajas, Tom Zidar, and Mert Bay. 2020. Advertising Incrementality Measurement using Controlled Geo-Experiments: The Universal App Campaign Case Study. (2020).

[5] Ron Berman. 2018. Beyond the last touch: Attribution in online advertising. *Marketing Science* 37, 5 (2018), 771–792.

[6] Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83, 1 (2015), 155–174.

[7] Prasad Chalasani, Ari Buchalter, Jaynth Thiagarajan, and Ezra Winston. 2017. Counterfactual-based Incrementality Measurement in a Digital Ad-Buying Platform. *arXiv preprint arXiv:1705.00634* (2017).

[8] 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* (Washington, DC, USA) *(KDD '10)*. ACM, New York, NY, USA, 7–16. https://doi.org/10.1145/1835804.1835809

[9] Brian Dalessandro, Rod Hook, Claudia Perlich, and Foster Provost. 2015. Evaluating and optimizing online advertising: Forget the click, but there are good proxies. *Big data* 3, 2 (2015), 90–102.

[10] Facebook. 2019. About Facebook Conversion Lift Studies. https://www.facebook.com/business/help/688346554927374

[11] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* 38, 2 (2019), 193–225.

[12] Daniel N Hill, Robert Moakler, Alan E Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. 2015. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1839–1847.

[13] Garrett A. Johnson, Randall A. Lewis, and Elmar I. Nubbemeyer. 2017. Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research* 54, 6 (2017), 867–884. https://doi.org/10.1509/jmr.15.0297 arXiv:https://doi.org/10.1509/jmr.15.0297

[14] Jouni Kerman, Peng Wang, and Jon Vaver. 2017. *Estimating Ad Effectiveness using Geo Experiments in a Time-Based Regression Framework.* Technical Report. Google.

[15] Pavel Kireyev, Koen Pauwels, and Sunil Gupta. 2016. Do display ads influence search? Attribution and dynamics in online advertising. *International Journal of Research in Marketing* 33, 3 (2016), 475–490.

[16] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge University Press.

[17] Randall A. Lewis and Justin M. Rao. 2015. The Unfavorable Economics of Measuring the Returns to Advertising *. *The Quarterly Journal of Economics* 130, 4 (2015), 1941–1973.

[18] Randall A. Lewis, Justin M. Rao, and David H. Reiley. 2011. Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/1963405.1963431

[19] Hongshuang (Alice) Li and P.K. Kannan. 2014. Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment. *Journal of Marketing Research* 51, 1 (2014), 40–56. https://doi.org/10.1509/jmr.13.0050 arXiv:https://doi.org/10.1509/jmr.13.0050

[20] Tesary Lin and Sanjog Misra. 2020. The Identity Fragmentation Bias. arXiv:2008.12849 [econ.EM]

[21] Donald B Rubin. 2005. Causal Inference Using Potential Outcomes. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331. https://doi.org/10.1198/016214504000001880 arXiv:https://doi.org/10.1198/016214504000001880

[22] Xuhui Shao and Lexin Li. 2011. Data-driven Multi-touch Attribution Models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (San Diego, California, USA) *(KDD '11)*. ACM, New York, NY, USA, 258–264. https://doi.org/10.1145/2020408.2020453

[23] Raghav Singal, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar. 2019. Shapley Meets Uniform: An Axiomatic Framework for Attribution in Online Advertising. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1713–1723. https://doi.org/10.1145/3308558.3313731

[24] Gabriele Tolomei, Mounia Lalmas, Ayman Farahat, and Andrew Haines. 2019. You must have clicked on this ad by mistake! Data-driven identification of accidental clicks on mobile ads with applications to advertiser cost discounting and click-through rate prediction. *International Journal of Data Science and Analytics* 7, 1 (2019), 53–66.

[25] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. 2015. Robust Tree-based Causal Inference for Complex Ad Effectiveness Analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM '15)*. ACM, New York, NY, USA, 67–76. https://doi.org/10.1145/2684822.2685294