Near Impressions for Observational Causal Ad Impact

Stephanie Sapp, Jon Vaver, Jon Schuringa, Steven Dropsho

Google Inc.

Abstract

Advertisers often estimate the performance of their online advertising by either running randomized experiments, or applying models to observational data. While randomized experiments are the gold standard of measurement, their cost and complexity often lead advertisers to rely instead on observational methods, such as attribution models. A previous paper demonstrated the limitations of attribution models, as well as information issues that limit their performance [1]. This paper introduces "near impressions", an additional source of observational data that can be used to estimate causal ad impact without experiments. We use both simulated and real experiments to demonstrate that near impressions greatly improve our ability to accurately measure the true value generated by ads.

1 Introduction

Advertisers need to understand the causal impact of their advertising in order to make sound marketing decisions. Properly randomized and controlled experiments are the most accurate way to measure the impact of advertising. For example, Google's ghost ad methodology is a technology that identifies control group users in a cookie level experiment who would have been exposed to an ad in the absence of the experiment [2]. However, the sophistication of this technology adds complexity to the ad serving systems. In these experiments, user lists must be maintained for each advertiser, and additional systems are required to simulate auction results for control cookies. Further, these experiments are often costly, since impressions and revenue are lost by actively withholding ads that would have otherwise been served. Few ad serving systems currently offer ghost ad experimentation, and its complexity could limit broader adoption.

Intent-to-treat (ITT) experiments are another unbiased alternative. This type of experiment compares outcomes for all users across the test and control groups. Specifically, ITT estimators measure the outcome for each user assigned to the test group, regardless of whether the user was exposed to the advertising being measured. Similarly, the outcome for each control group user is measured, regardless of whether the user would have been exposed to this advertising. Examples of ITT include user level [3] and geo level [4] experiments. A major drawback to ITT experiments is that their estimates have relatively high variance due to the lack of exposure tracking. This greater variability makes it much more difficult to detect significant causal impact from treatment. Applied to advertising, this means that ITT experiments typically require large campaign changes in order to detect lift.

Public Service Announcement (PSA) tests are a third common type of experiment. PSA tests aim to identify the subset of control group users who would have been exposed, had they been assigned to the test arm, by serving PSA advertisements to users within the control arm [5]. However, this approach has several disadvantages.

First, PSA tests are costly, since advertisers must pay for the PSA ads served to users assigned to the control group. Second, complexity is introduced by the need to differentiate ad serving between test and control users. In particular, a mechanism is needed to randomly serve the PSA and non-PSA ads. Finally, PSA tests with modern ad servers can yield biased estimates, since ad serving systems use optimization to deliver those ads that generate the most clicks or conversions. This optimization causes PSA tests to deviate from the conditions needed for a valid experiment, by disrupting the comparability of the test and control groups.

Due to complexity, cost, and limitations, experiments are often impractical or impossible for advertisers to rely upon as a primary means of measurement. More typically, advertisers use experiments in a one-off capacity. Instead, advertisers use observational methods to measure and monitor the performance of their advertising. Attribution models are a common observational method used to accomplish this task for online advertising. Attribution models allocate credit for conversions in observational user-level path data to the marketing events observed prior to each conversion. Observational methods are easy to apply and offer the promise of always-on measurement. However, typical formulations of these models can yield misleading estimates of advertising effectiveness [1].

Improved modeling, such as the Upstream Data-Driven Attribution (UDDA) model described in [1], can significantly improve the accuracy of attribution estimates. However, limitations in information available in the attribution data were also shown to be another factor that limits the accuracy of attribution models. First, attribution models have no visibility to users who have no observed interactions with the advertiser. Second, there can be unobserved dissimilarity in the browsing behavior of exposed and unexposed users, since many unexposed users were not active on websites serving the ads being measured. Third, ads are often targeted towards users who behave differently than untargeted users. All three issues lead to bias in the comparison groups of unexposed users who serve as pseudo-controls for the exposed users.

This paper introduces *near impressions*, a source of observational data that can be used in attribution models to construct naturally occurring pseudo-control groups that enable us to recover causal ad impact in situations where these models currently fail. Compared to traditional experiments, near impressions are more scalable, less costly, have simplified infrastructure requirements, and enable ongoing monitoring of ad effectiveness.

2 Near Impressions

2.1 Description

Near impressions are observational, user-level events that identify instances in which a user activity generated an ad serving opportunity that matched an advertiser's targeting criteria, but didn't result in the user being exposed to the advertiser's ad. Examples include instances in which an ad loses an auction, an advertiser is out of budget, or an ad system is overbooked. Near impressions are very powerful because they allow us to differentiate among unexposed users: we can identify the subset of unexposed users who were both active and targeted by the advertiser. This is important because this group of users may behave similarly to the exposed users, and very differently than users who were not active or targeted by the advertiser. These unexposed users with near impressions serve as weighted observational pseudo-controls, as described in detail in Section 2.2. They allow us to better estimate causal ad impact without experiments.

The manifestation of near impressions is flexible, and can vary by ad channel. For example, near impressions can include served but unviewable ads, or instances of an ad that nearly missed being delivered by the ad server. In a single-

winner auction environment, a near impression event based on a lost auction occurs when an ad participated in, and was competitive in, the auction, but did not win. These near impression events are identified through auction bid data. For reservation-based ads, near impression events can be identified as eligible but unrealized ad impression opportunities within the targeted reservation inventory. These lost opportunities could be caused, for example, by the advertiser being over budget, or by another eligible advertiser's impression being shown.

This paper uses advertising in a single-winner ad format auction as an example of using of near impressions to estimate ad effectiveness. For these near impressions, the nearness criteria can be viewed as a bias-variance tuning parameter. Including only those losing auctions that were very near to winning is best for removing bias, while including all losing auctions generates more near impressions and more opportunities for counterfactual comparisons.

2.2 Ad Impact Estimation: UDDA-NI

The UDDA with Near Impressions (UDDA-NI) algorithm is a simple modification to UDDA that incorporates near impression information. UDDA aggregates users with the same sequence of events upstream from the index of ad exposure, and matches them with unexposed user paths which have the same upstream sequence of events, but no ad exposure at the exposure index. UDDA-NI operates similarly to UDDA, but further restricts the unexposed user set to those paths that also have a near impression at the exposure index. That is, UDDA-NI still matches users based on their upstream paths, and in addition matches ad exposure events with near impression events. By introducing this additional matching of impressions with near impressions, we obtain a control group that matches the test group better than one that includes all unexposed users.

Figure 1 illustrates the matching procedure used in the UDDA-NI algorithm. UDDA-NI assigns credit to an ad exposure within a user path by comparing the conversion rate of user paths with a common sequence of marketing events upstream from an ad exposure to the conversion rate of paths that have the same sequence of events upstream from a near impression event instead of an ad exposure. The difference in conversion rates between the exposed and near impression paths estimates the ad's effectiveness within this set of exposed paths. Total ad effectiveness within this exposed set is found by aggregating ad effectiveness estimates for all exposed users in the set. These credits are aggregated across all unique upstream paths among the exposed users to calculate an overall ad effectiveness.

We provide a formal description of the UDDA-NI algorithm below. Note that these steps are analogous to the UDDA algorithm described in [1]; the primary modification occurs at Step 3b, where we impose near impression matching at the exposure index for the set of unexposed users.

Using the same notation as [1], let user i have an ordered path of observed events denoted by the vector $X^i = (X_1^i, \ldots, X_{L(i)}^i)$, with length $L(i)$. Let A denote an ad impression exposure of the type currently being analyzed, let N denote a near impression of the type currently being analyzed, and let C denote a conversion event. For the index m, let $U_m^i = (X_1^i, \ldots, X_{m-1}^i)$ denote the upstream path that includes the first $m-1$ events in a user path, i.e. the sequence of events that occurred prior to index m in the path. The UDDA-NI algorithm then proceeds as follows:

1. Classify all user paths as either containing or not containing an ad impression A:
   (a) For all exposed paths, i.e. $\exists A \in X^i$, set exposure indicator $T_i = 1$.
   (b) For all other paths, which are unexposed, set $T_i = 0$.

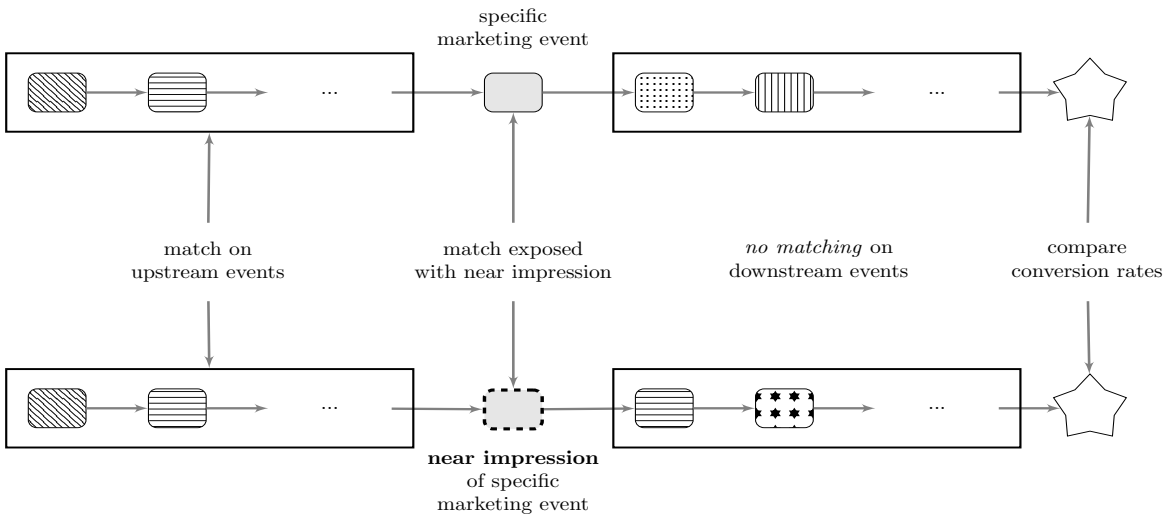


Figure 1: Illustration of UDDA-NI algorithm. Users with the same observed sequence of events prior to the index of ad exposure are aggregated. A set of unexposed users with a near impression in the exposure index, and the same upstream sequence of events is also aggregated. The difference in conversion rates between the exposed and near impression paths estimates the ad effectiveness for the set of exposed paths. Overall ad effectiveness is found by aggregating estimates across all unique upstream paths among the exposed users.

2. For each user path i with $T_i = 1$:

   (a) Let m_i denote the index of the first occurrence of A in the path. That is, $X_{m_i} = A$, and no previous event in the path equals A.

   (b) Record the upstream path of user i as the sequence of events prior to this exposure index: $U^i_{m_i}$.

   (c) Calculate the number of downstream conversions that occur after index m_i as: $C_i = \sum_{j=m_i+1}^{L(i)} I(X^i_j = C)$.

3. For each unique upstream path u_j from Step 2b:

   (a) Let $n_j(T = 1)$ denote the number of users in the exposed group with upstream path u_j. Calculate the average conversion rate among these users as: $\bar{c}_j(T = 1) = \frac{1}{n_j(T=1)} \sum_{i=1}^{n_j(T=1)} C_i$.

   (b) Find all unexposed users i with a near impression at index m_j, i.e. $X_{m_i} = N$, and upstream path u_j from index m_j, i.e. $U^i_{m_j} = u_j$. Calculate the number of downstream conversions C_i after index m_j for each of these users analogous to Step 2c. Calculate the average conversion rate $\bar{c}_j(T = 0)$ among these users analogous to Step 3a.

   (c) Estimate the incremental conversion rate among these users as the difference in conversion rate among exposed versus near impression users: $\hat{r}_j = \bar{c}_j(T = 1) - \bar{c}_j(T = 0)$.

   (d) Estimate the number of incremental conversions among these users as: $\hat{r}_j \cdot n_j(T = 1)$.

4. Aggregate the estimated incremental number of conversions among all exposed user paths by aggregating over all unique upstream paths: $\hat{\theta} = \sum_{u_j} \hat{r}_j \cdot n_j(T = 1)$.

2.3 Regression Discontinuity

Near impressions from lost auctions[1] are conceptually similar to regression discontinuity designs. Regression discontinuity designs provide an observational method for estimating causal impact of an intervention when treatment assignment is determined by a common exogenous threshold [6], and they have been applied in advertising research to study position effects in search ads [7]. A classic application of regression discontinuity designs is the evaluation of scholarship programs, as in [8]. In this setting, the causal impact of receiving the scholarship is estimated by comparing outcomes among individuals with test scores very close, both above and below, to the threshold for awarding scholarship. Since these individuals are presumably similar, differences between their later outcomes are likely due to receiving the scholarship or not.

While similar in flavor, we highlight some key differences between near impressions from lost auctions and regression discontinuity designs. First, auctions do not have a common, fixed winning threshold. Instead, the threshold for winning an auction varies across auctions, and we aggregate losing bids both within each upstream path, and then across all unique upstream paths, as described in Section 2.2. The effect of this aggregation is to estimate the average causal impact of winning an auction, averaged over a mix of different observed thresholds, and is likely representative of the similar mixture of thresholds in future auctions.

Second, in the conversion rate comparison, we include all winning auctions, rather than only winners near the threshold. In the scholarship program evaluation setting, this comparison would result in overestimation bias, since individuals with test scores significantly above the threshold are more likely to have successful outcomes, compared to those with scores near the threshold. An auction setting is less likely to facilitate a similar overestimation bias due to competition within the market dynamic, since auctions with different thresholds are grouped together within the test and control groups. These pooled sets of data are used to estimate the average causal impact of winning the auction across a range of thresholds, and it is unlikely that advertisers will win only the most effective auctions while losing the less effective ones. It is difficult for a given advertiser to win only the most effective auctions, while losing the lesser opportunities. Further, matching on upstream paths of observed activities prior to exposure is designed to mitigate any potential bias, similar to the inclusion of covariates to reduce bias in a regression discontinuity design. In the performance evaluation results that follow in Section 3, we do not observe significant bias with near impressions.

[1] Near impressions from lost auctions are the focus of this paper, but near impressions are applicable beyond auctions (for example, in reserved ad buy situations). See Section 2.1 for further details.

3 Performance Evaluation

We evaluated the performance of near impressions using both simulated and real user-level path data. In both evaluations, an experimental ground truth is available, which allows us to compare each observational estimate using near impressions to the corresponding experimentally determined result.

3.1 Simulations

The Digital Advertising System Simulation (DASS) [9] is a simulation framework which models online advertising and its impact on user behavior using an extended, non-stationary Markov model. The simulation consists of a user activity path model for user browsing behavior in the absence of advertising, an ad serving model for the process by which users are exposed to advertising events, and an ad impact model for how exposure to advertising impacts downstream user behavior.

Since the original version of DASS described in [9] does not generate near impression events, we introduced a modification to generate lost ad serving opportunity events. Specifically, we modified DASS to record lost ad serving opportunity event types as instances in which a user met the campaign's serving criteria, but the ad was not served due to share of voice constraints. That is, the user's activity state was one on which the ad type was eligible to be served, the user's impressibility to the ad type met the ad's minimum threshold, and the user's number of previous exposures to the ad was below the ad's frequency cap, but the ad was not served to the user because the random share of voice Bernoulli draw indicated that the ad should not be shown.

Four display ad simulation scenarios generated by DASS were provided in [1] to illustrate the capabilities and limitations of the UDDA algorithm. In this section, we apply near impressions via the UDDA-NI algorithm to the same four scenarios to demonstrate how near impressions can be used to address the three key challenges that limit the performance of UDDA without near impressions[2]. For each scenario, we compare the results of UDDA-NI to the UDDA algorithm with visibility to censored user paths (UDDA-VCU) but without visibility to near impressions. In real data, attribution models do not have visibility to users who have no observed interactions with the advertiser; these user paths are systematically censored. Visibility to censored paths is possible in the simulated environment of DASS, but is not possible in practice. That is, UDDA-VCU is included in this paper for theoretical performance comparison only, and is not a viable approach for real path data[3].

[2]A complete description of the parameterization of these four simulation scenarios is available in the Appendix of [1].

[3]The performance of UDDA *without* visibility to censored users (and without visibility to near impressions) is generally much worse than UDDA-VCU.

3.1.1 Systematically Censored Users

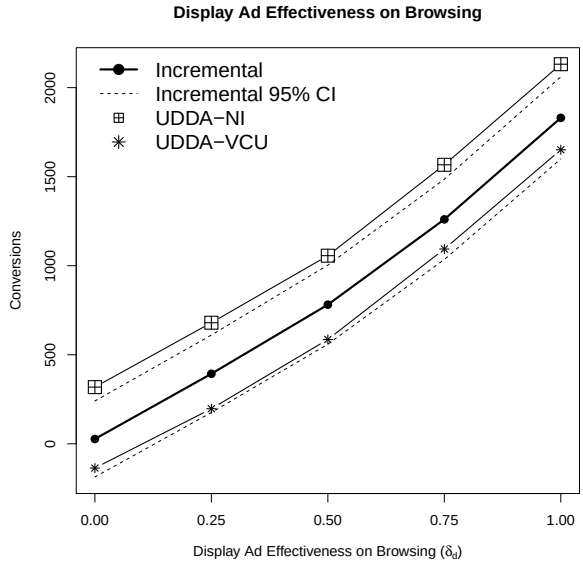
The first two scenarios illustrate how near impressions can address the challenge introduced by the systematic censoring of users who have no observed interactions with the advertiser. We show results for two possible ways in which display ads might impact user behavior. Figure 2(a) shows results for the scenario in which display ads impact user browsing behavior: increasing the user's likelihood of performing a related branded or generic search, or visiting the advertiser's website. Figure 2(b) shows results for the scenario in which display ads directly impact user conversion probability if the user happens to visit the advertiser's website, but do not change the user's downstream browsing behavior in any other way.

In these scenarios, UDDA-VCU is able to accurately recover the true number of incremental conversions generated by the display ads, and UDDA-NI achieves comparable performance. Most importantly, UDDA-NI is a feasible algorithm for real path data, while UDDA-VCU is not.

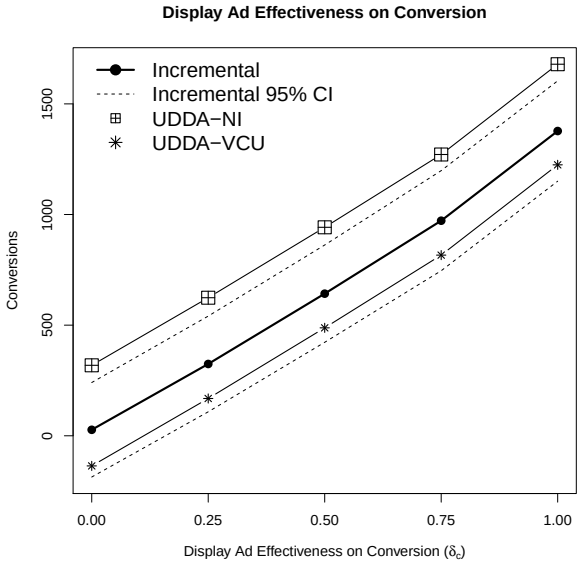
3.1.2 User Browsing Dissimilarity

The next scenario models dissimilarity between the browsing behavior of exposed versus unexposed users. That is, all users exposed to a given ad type visited a website on which that ad was eligible to be served, but this is not the case for all unexposed users. Specifically, not all unexposed users had the opportunity to be served the ad. Results for this scenario are presented in Figure 2(c).

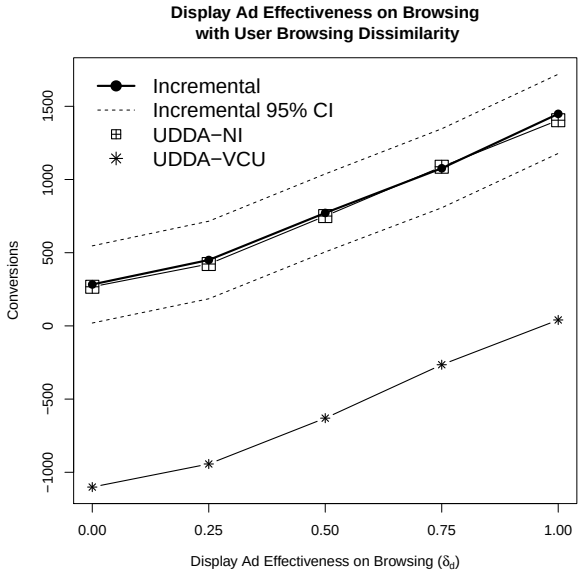
UDDA-VCU is not able to capture the true value generated by display ads in this scenario, even with visibility to censored users, since the unexposed user set in this scenario includes users who never browsed to an activity state on which the ads were eligible to be served. In contrast, by incorporating near impression information, the



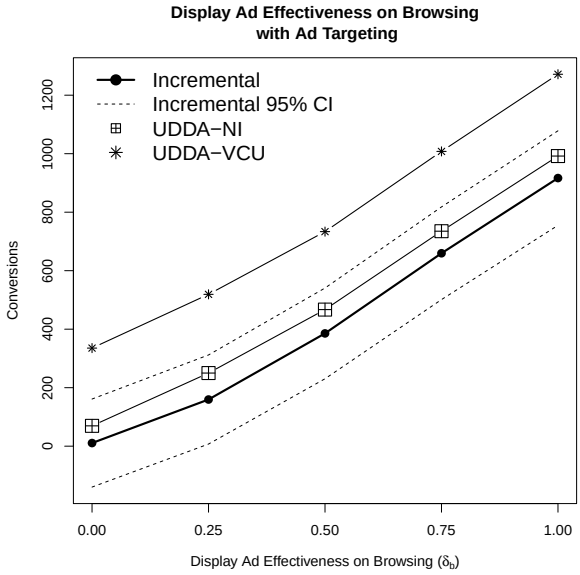
(a) Simulations that vary the level of display ad effectiveness in changing user browsing behavior.



(b) Simulations that vary the level of display ad effectiveness in directly changing user conversion rate.



(c) Simulations that vary the level of display ad effectiveness in changing user browsing behavior, modified to include user browsing dissimilarity. Some unexposed users never had opportunity to be served a display ad.



(d) Simulations that vary the level of display ad effectiveness in changing user browsing behavior, modified to include ad targeting. Some unexposed users do not meet the ad's targeting criteria.

Figure 2: Performance comparison on simulated datasets of near impressions via the UDDA-NI algorithm, versus the idealized but infeasible UDDA-VCU. Near impressions are an effective addition to the UDDA algorithm that provide a viable way to recover incremental conversions generated by ads, even in situations in which the incorporation of visibility to censored users is unsuccessful.

UDDA-NI algorithm is able to accurately estimate the true number of incremental conversions generated by the display ads. In this case, the near impressions allow us to identify unexposed users who did browse to an activity state on which the ads were served.

3.1.3 Ad Targeting

The final scenario simulates ad targeting, with the display ads targeted towards users who are inherently more likely to convert. Users who do not meet the ad's targeting criteria are less likely to convert, on average. Figure 2(d) shows results for this scenario.

Visibility to censored users is again not sufficient, since UDDA-VCU is not able to accurately estimate the true number of incremental conversions from the display ads in this scenario. Here, all unexposed users browsed to an activity state on which the ads were eligible to be served, but the unexposed user set also includes users who were not targeted by the ad. By adding near impression information, the UDDA-NI algorithm can again recover the true value of the display ads. The near impressions make it possible to pinpoint unexposed users who did meet the ad's targeting criteria.

3.2 Real Ad Experiments

Google's ghost ad technology is a framework that enables advertisers to run experiments to measure the impact of their display and video ad campaigns [2]. Ghost ad experiments randomly assign users to either a test or control arm. Users within the test group are eligible to be served ad impressions from the advertiser, while users within the control group have the advertiser's ads withheld. Within the test group, exposed users are identified as those who are served one or more ad impressions. Ghost ads allow us to identify a comparable set of users, those users who would have been exposed if not for the experiment, within the control group. The ground truth incremental conversion lift for each experiment is calculated based on the difference between conversion rates for the exposed test group users versus the would-have-been-exposed control group users.

We evaluated the performance of near impressions by comparing estimates from the UDDA-NI algorithm to experimental measurements of lift from 178 display and video ad experiments. All experiments measured an online conversion outcome, but different conversion types were measured across the experiments. For example, conversion outcomes ranged across engagement type off-site conversions (such as subscribing to advertiser's social media channel, or watching more of the advertiser's videos on YouTube) to advertiser-defined on-site conversions (such as traditional e-commerce conversions, or reaching a particular page on the advertiser's website).

Each study's advertiser-specific impressions, near impressions[4], and conversions among the test arm users were made available to the UDDA-NI algorithm, since only the test group contains the observational data that is typically available to an observational model. That is, we did not use any data dependent on the experiment to compute the observational estimates. This allows us to evaluate methodologies in the same environment as would be used in practice, that is, one in which an experiment is not being run. As a result, UDDA-NI's upstream path matching was limited to only prior conversions and prior near impressions.

For comparison, we also computed results for a standard attribution model[5], without access to near impressions. Figure 3 shows an exam-

[4]Lost auctions have a reported auction score. Various relative and absolute thresholds were used to constrain the lost auctions that were deemed to be near impressions. In this application, algorithm performance was not strongly dependent on the choice of threshold.

[5]Results shown are independent of the model used, since typical attribution models are subject to reporting constraints that cause them to assign the same aggregate credit when the data contains only one event type.

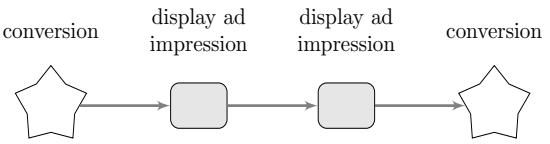


Figure 3: Example user path without access to near impression events from a display ad experiment. Only display ad impression and conversion events are observable. All standard attribution models will ignore the first conversion, since it occurs prior to any other observed events, and assign credit for the second conversion to the display ad. Note that any potential split between the first and second display ad is irrelevant for this example, since the total credit assigned to this type of display ad remains the same.

ple user path without near impressions from a display ad experiment. Note that without near impressions, only impressions and conversions within the test group users were made available to the standard attribution algorithm. As a result, any typical attribution model will attribute all conversions following an impression, or sequence of impressions, to that one type of impression. This behavior is caused by the requirement that attribution credits assigned within each user path must sum to the total number of conversions that occur after the first ad impression. That is, the impression is credited for all conversions following it (or them), since no other event types are available to share the credit. This is true for all current common attribution models, both position rule-based models and data-driven models.

Since the lift point estimates and uncertainties vary greatly across the 178 experiments, we evaluated the performance of the observational methods via a standardized evaluation metric. For a given study k, let δ_k and σ_k denote the experimentally-determined incremental conversion lift point estimate and standard error for that experiment, and let $\hat{\theta}_k$ denote the UDDA-NI estimate of incremental conversions generated by the advertising. We then use the following evaluation metric ϵ_k to measure the performance of UDDA-NI for experiment k:

$$\epsilon_k = \frac{\hat{\theta}_k - \delta_k}{\sigma_k} \quad (1)$$

An analogous evaluation metric is used to measure the performance of the standard attribution model estimates.

This ϵ metric measures the distance between the observational method's estimate and the experimental result, standardized by the experiment's uncertainty. Standardization allows us to compare across a large number of studies, by moving to a common scale. A metric value of zero is the best possible value, as this means that the observational and experimental estimates are the same. However, by randomness alone, these results will not agree exactly, so we need to consider the distributions of the metric values across experiments.

We highlight the percentage of studies in which each observational method's estimates are within the 95% confidence interval of the experimental measurement. Method estimates that satisfy $-1.96 < \epsilon < 1.96$ are within a 95% confidence interval for the experimental result. By randomness alone, we expect 5% of the results to lie outside of a 95% confidence interval. So, 95% of results within the 95% confidence interval is the gold standard[6].

Results are shown in Figure 4. Figure 4(a) compares near impressions to standard attribution model performance for each study. Each dot shows the value of $|\epsilon_k|$, for near impressions on the x-axis, and for standard attribution on the y-axis. Points lying on the 45° line indicate studies where the methods have the same absolute metric value. Near impressions have lower error for points falling above this line, and higher error for points below the line. Near impressions achieve lower or equal error for 89% of the experiments. Note that when the methods have equal absolute error, this error tends to be low, since most points on the 45° line are clustered near zero,

Figure 4(b) shows the distribution of ϵ values per

[6]Note that we are implicitly assuming that the observational method has no random error here. A more formal evaluation would require incorporating its variance and covariance.

(a) Paired comparison of method performance for each study. Absolute number of experimental standard errors is shown for near impressions on the x-axis, and for standard attribution model on the y-axis. The 45° line marks studies where methods have equal absolute error. Points above the line show studies where near impressions have lower absolute error.

(b) Boxplots of distribution of number of experimental standard errors per method. Median number of experimental standard errors per method is labeled, boxplot boundaries are 25th and 75th quantiles, boxplot whiskers are 10th and 90th quantiles, and horizontal dashed lines represent a 95% confidence interval for the experimental result.

Figure 4: Performance comparison of near impressions via the UDDA-NI algorithm, versus standard attribution model estimates, for 178 display and video ad studies across a variety of online conversion types. Each point shows the value of Equation 1 for the corresponding observational method for one study. That is, each point shows the difference between the observational estimate and the experimental measurement of incremental conversions generated by the ads, standardized by the experimental uncertainty, for one study. Estimates using near impression information systematically outperform estimates that do not use this information.

method across the studies. The median metric value per method is labeled. The boundaries of each box are the method's 25th and 75th quantiles, and the whiskers are the method's 10th and 90th quantiles. The horizontal dashed lines across the entire plot at ± 1.96 represent a 95% confidence interval for the experimental result. In particular, any dot that is within these lines means that the method's estimate is within the 95% confidence interval for the experimental result. As mentioned above, a gold standard of performance is for 95% of the experiments to be within the 95% confidence interval.

Near impressions achieve a median standardized error of 0.2, while the standard attribution model has a median standardized error of 0.8. The near impressions based estimate is within the experimental 95% confidence interval for 88% of the experiments, compared to 75% of the experiments with the standard attribution model. Overall, the near impressions method produces estimates that are closer to the experimental result than the standard method, as the near impressions estimates cluster more closely around zero.

While near impressions produce promising performance, their estimates are not perfect. Since we evaluate performance compared to experiments, which are the gold standard of measurement, we don't expect any observational method to be flawless. For UDDA-NI, there may be residual differences between the exposed and near impression user groups. We conclude with some thoughts towards further improving performance.

First, exact upstream path matching can induce sparsity from users with uncommon upstream paths. This type of data sparsity could be better handled by generalizing the upstream matching procedure to group together users with similar, but not identical, upstream paths. Second, grouping all impressions (with the same upstream path) together, and matching them with all near impressions (having the same upstream path) may produce user groups that are not sufficiently comparable. Refining the impression and near impression grouping and matching, for example by requiring groups and matches to come from the same campaign, may yield more comparable user groups. Finally, the real user path data we used in this paper limited matching to only prior conversions and prior near impressions. Adding additional user behaviors and characteristics for upstream matching could reduce differences between the groups.

4 Concluding Remarks

In this paper, we introduce near impressions, an abundant type of observational data that can be used to identify the subset of unexposed users who were both active and targeted by an advertiser's ad. We describe how near impression events can be incorporated into the UDDA attribution algorithm. Finally, we illustrate promising performance using near impressions to estimate the incremental value generated by advertising using both real and simulated experimental results from single-winner ad formats.

We are continuing research on applying near impressions to measure advertising impact more broadly. For example, we are working to extend near impressions to multi-winner ad formats (such as search ads), incorporate near impressions into digital attribution modeling, and collect currently unavailable conversion outcomes for near impression users (such as survey responses).

Near impressions have the potential to make observational measurement more useful and trustworthy. They enable measurement for ads where experiments are not currently available. For ads in which one-off experiments are possible, they open the door for always-on measurement. The results in this paper show that near impressions based estimates can be similar to experimental measurements, making it possible for advertisers to supplement experiments and measure ad impact at greater scale.

Acknowledgments

The authors wish to thank the many colleagues from Google whose contributions enhanced this paper, with special thanks to Art Owen for highlighting connections to and differences from regression discontinuity, and Baokui Yang and Bill Halpin for supporting our display ads analysis.

References

[1] Stephanie Sapp, Jon Vaver. "Toward Improving Digital Attribution Model Accuracy" *Google Inc.*, 2016. https://research.google.com/pubs/pub45766.html

[2] Garrett A. Johnson, Randall A. Lewis, Elmar I. Nubbemeyer. "Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness" *Journal of Marketing Research (In-Press)*, 2017.

[3] Thomas Blake, Chris Nosko, Steven Tadelis. "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment" *Econometrica 83(1):155–174*, 2015.

[4] Jon Vaver, Jim Koehler. "Measuring Ad Effectiveness Using Geo Experiments" *Google Inc.*, 2011. https://research.google.com/pubs/pub38355.html

[5] Judy Morimoto. "Tips for Running PSA (Test and Control) Campaigns" *3Q Digital*, 2013. https://3qdigital.com/display/tips-running-psa-test-control-campaigns/

[6] Guido W. Imbens, Thomas Lemieux. "Regression Discontinuity Designs: A Guide to Practice" *Journal of Econometrics 142(2):615–635*, 2008.

[7] Sridhar Narayanan, Kirthi Kalyanam. "Position Effects in Search Advertising and Their Moderators: A Regression Discontinuity Approach" *Marketing Science 34(3):388–407*, 2015.

[8] Donald L. Thistlethwaite, Donald T. Campbell. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment" *Journal of Educational Psychology 51(6):309–317*, 1960.

[9] Stephanie Sapp, Jon Vaver, Minghui Shi, Neil Bathia. "DASS: Digital Advertising System Simulation" *Google Inc.*, 2016. http://research.google.com/pubs/pub45331.html