Attribution Model Evaluation

Kyra Singh, Jon Vaver, Richard Little, Rachel Fan

Google LLC

Abstract

Many advertisers rely on attribution to make a variety of tactical and strategic marketing decisions, and there is no shortage of attribution models for advertisers to consider. In the end, most advertisers choose an attribution model based on their preconceived notions about how attribution credit should be allocated. A misguided selection can lead an advertiser to use erroneous information in making marketing decisions. In this paper, we address this issue by identifying a well-defined objective for attribution modeling and proposing a systematic approach for evaluating and comparing attribution model performance using simulation. Following this process also leads to a better understanding of the conditions under which attribution models are able to provide useful and reliable information for advertisers.

1 Introduction

Advertisers are interested in understanding and measuring the impact that campaigns have on consumer behavior. In the digital world, when a user converts (i.e. makes a purchase, signs up for a mailing list, etc.) it is useful to know how that user arrived on the advertiser's website (e.g., paid search click, organic click, direct navigation, email link, etc.). This information provides insight into consumer behavior and preferences, which can be interpreted and translated into tactical and strategic marketing actions. This desire to know where users come from before converting is the origin of digital attribution. The earliest and simplest form of digital attribution is the last interaction, or last event, model [Help, 2017b]. This model gives "credit" to the last event in a user's browsing path prior to conversion. This last event information is readily available in the referring URL and the output of this model is easy to understand.

With advancements in digital technology, largely the widespread use of cookies and ad tagging, it became possible to account for website visits and marketing interventions that are further upstream from a conversion. These new data sources made it possible to generalize the notion of crediting user activity to conversions. However, the question that results is *how* should this credit be assigned across multiple events?

Many different approaches have been developed to address this question. In addition to the last interaction model, there are other "rules-based" models that assign fractional credit according to weights that are determined by the position and number of events in the user path. Two examples are the first interaction and linear attribution models [Help, 2017b]. Data-driven attribution (DDA) models are more sophisticated and distribute credit across multiple touch points by considering converting and non-converting paths to model the probability of conversion across different paths. Upstream DDA allocates credit based on the probability of conversion, as determined by matching upstream events of users exposed and unexposed to an ad event [Sapp and Vaver, 2016]. Models introduced by Shao and Li [2011] and Dalessandro et al. [2012] use logistic regression to predict the occurrence of conversions following ad events, and Li and Kannan [2014] describes an attribution model that works within a Bayesian framework.

These are just a few examples of the attribution models described in the literature, and it is likely that there are many others used by advertisers and third party measurement providers that have not been described publically.

Ultimately, advertisers must choose among this abundance of attribution models. This choice is typically made without tangible evidence that one model will objectively outperform another in the most important and relevant situations. A formal process for evaluating attribution models is needed to fill this gap. Having such a process has additional benefits. It can identify the situations in which models have differentiated performance, help set realistic expectations for model performance, and indicate where models need improvement. This paper describes a process for evaluating attribution models that relies on simulation.

This paper is organized as follows. Section 2 provides a framework of current attribution expectations and assumptions. Section 3 defines the primary attribution objective used in this paper. In Section 4, the simulation process used to evaluate attribution models is presented, and Section 5 describes the different ways in which advertising can impact user behavior within the simulation. Section 6 contains a categorization of advertising conditions, metrics for scoring, and a discussion of evaluation results. A brief summary and conclusion are provided in Section 7.

2 Attribution Expectations

The proliferation of attribution models may lead advertisers to place a great deal of focus on attribution model selection. However, this decision cannot be made without taking other considerations into account. The quality of information generated by an attribution model also depends on reporting constraints, the type of user-level event data that is available (i.e., the data scope), and the identified modeling objective (as discussed in Section 3).

The last interaction model associates a single unit of credit with every conversion that has a trackable upstream event. The number of credits assigned within each path equals the number of conversions within that path, and the total credits assigned to user-level events equals the total number of "attributable" conversions. We refer to this property as the "last event accounting principle." Due to its familiarity, convenience, and interpretability, this principle has been preserved in the development of newer attribution models. However, this principle does not align with the objective of measuring ad effectiveness. For example, in user paths where there is only a single paid ad event followed by a conversion, the accounting principle requires that full credit be assigned to the paid ad, even if the ad is completely ineffective. In these cases, the advertiser will be misled about ad effectiveness since credit is assigned to an event that did not actually impact the outcome.

A separate issue, which compounds the problems caused by conforming to last event accounting, is that not all marketing events are available in the attribution modeling process. These events are "out-of-scope" for attribution modeling. Offline advertising is one source of out-of-scope advertising, which includes television advertising. However, even digital ads can be fully, or partially, out-of-scope. Clicks are more easily tracked than impressions, so paid ad clicks might be in-scope, while the associated impressions are out-of-scope. Assignment of credit to ineffective events can also occur when there are multiple events in a path because the effective advertising events are out-of-scope. For example, advertisers may not have complete information regarding exposures to advertising events on a third party website. Reporting is limited to in-scope events, even though out-of-scope events may have impacted user behavior. This data completeness issue can lead to incorrect allocation of credit to in-scope advertising events.

Although attribution has the stated goal of allocating conversion credit, this goal is not consistent with the goal of understanding advertising effectiveness. Beyond this mismatch, the goal of allocating conversion credit does not have an objective truth, and attribution model evaluation and comparison requires this information. In order to evaluate and compare the effectiveness of attribution models, it is necessary to identify a well-defined attribution objective.

3 Causality as an Attribution Objective

A scientific evaluation of attribution model performance must begin with a clearly stated objective. Advertisers need to understand the effect of ads on consumer behavior with the ultimate goal of quantifying the impact on conversion volume. Therefore, attribution should be considered a causal estimation problem. Causal measurement makes it possible to say how effective advertising is at changing user behavior. So, we have made it the cornerstone for evaluating attribution models [Kelly et al., 2018]. The optimal way to measure causal impact is to conduct a fully randomized controlled experiment and compare the outcomes of treated versus untreated subjects [Rubin, 1974]. In advertising, the treatment is ad exposure and the subjects are the users.

Different experiments are needed for different causal measurement objectives. Estimating the number of incremental conversions (IC) generated by each ad channel, the marginal IC (mIC) rate of each ad channel, and the incremental conversions generated by each individual ad event are all objectives of potential interest to advertisers. For IC, the objective is to measure the difference between the number of conversions with an ad channel present versus not present (the ad channel on versus the ad channel off). The evaluation process described in this paper concentrates on this IC objective, although a similar process is applicable for other causal measurement objectives.

Conducting a real world experiment is the ideal mechanism for generating causal measurements that can be used to evaluate attribution models. However, this is a costly and impractical approach due to the potentially large number of ad channels, the complexity of ad campaigns, and the reluctance of advertisers to forego the opportunity to serve ads [Chan et al., 2010]. Evaluating attribution models requires an alternative approach. We leverage the DASS tool introduced in Sapp et al. [2016] to accomplish this goal.

4 Simulation

The Digital Advertising System Simulation (DASS) is a flexible framework developed for modeling advertising and its impact on user behavior Sapp et al. [2016]. DASS has the ability to generate sets of path data under a wide variety of marketing conditions. A Markov process models user behavior in the absence of advertising, and ads are injected into the user's browsing stream. These ads have the ability to modify the transition matrix of the Markov process thereby affecting user behavior, which can increase the probability of conversion. Through the specification of simulation parameters, the system provides control of browsing activities, user characteristics, the mix of users in a simulation, types of advertising, how ads are served to users, the impact of ads, and more.

Events in the original DASS model do not include time stamps, and therefore it has a limited ability to model ad impact that changes over time. Since these situations are of interest for this paper, and other applications, DASS has been extended to address this shortcoming. This time-based version of DASS is described in Appendix A. It closely follows the original DASS framework with additional parameters that generalize the ad impact model.

It is possible to run "virtual experiments" with both versions of DASS. For each ad channel b_j, where $j = 1, \ldots, J$ are the channels included in the simulation, an experiment is run. In each experiment, two sets of path data are generated; one with all ad channels $1, \ldots, J$ on and the other with a single ad channel, k, turned off. The difference between the number of conversions generated by these two sets of paths is the incremental value of the k^{th} ad channel. The results of such experiments provide the ground truth needed to evaluate and compare attribution models. It also aligns the evaluation with the causal objective described above. However, the key to the evaluation is the specification of the underlying simulation parameters, which determine the marketing conditions under which the attribution models are being asked to perform.

5 Types of Ad Effectiveness

Advertising can impact user behavior in various ways. Within the simulator, we can specify both the mechanism and strength of ad impact on user behavior. Figure 1 illustrates different ways that a search ad can impact user behavior.

The most direct impact that a search ad impression can have on user behavior is to encourage the user to visit the advertiser's website directly via a paid click, which provides the opportunity to make a purchase (i.e., to generate a conversion). The magnitude of this impact is controlled via the Click Through Rate (CTR) for an ad. An additional click-related factor is the bounce rate. This indicates the probability that the user finds the advertiser's website to be irrelevant, or accidentally clicks, and immediately resumes their previous browsing activity without impact.

Alternatively, a search ad may have a less direct impact on user behavior. See the "impression effect" and "click effect" connectors in Figure 1. For example, search ad impressions and/or search ad clicks, may encourage users to change their downstream browsing behavior. At a later time, exposed users may be more likely to do another search, perform a branded search, or visit the advertiser's website directly. These behavioral changes are realized in DASS through modifications of the user transition matrix, which controls user browsing behavior. Figure 1 also reflects the possibility that a search impression could impact the rate at which a user will convert, conditioned on a visit to the advertiser's website, without otherwise impacting the user's downstream browsing behavior. This impact is analogous to the impact that brand advertising can have on offline purchases. A brand ad may not drive a user to the store, but it may help with brand selection once the user is in the store.

Downstream impact on user behavior can be temporary, diminishing very quickly after ad exposure, or persistent, permanently impacting all future browsing behavior. When a user is exposed to an ad, components of that user's transition matrix are scaled to reflect ad impact and, as time passes, the transition matrix will revert towards its initial specification. The *persistency of ad impact* (i.e., the rate of reversion) is controlled by separate parameters for impressions and clicks. These parameters are another way that ad effectiveness can be modulated in the simulation.

Finally, ads may be preferentially served to a specified set of users. That is, DASS allows for ad targeting, and each type of ad can have a different type or magnitude of impact across users. So, many combinations of ad channels, types of ad impact, magnitudes of ad impact, and audiences can be considered in an individual DASS scenario. Each scenario represents a different challenge for an attribution model, and we would like to know how well an attribution model estimates the causal objective under the advertising situation posed by the scenario. More interesting still is an assessment of an attribution model's ability to estimate a causal objective across a wide range of reasonable and informative DASS scenarios. This exercise is most useful when scenarios are created and organized with the goal of shining a light on the fundamental capabilities and limitations of an attribution model. This organization of these scenarios is discussed next.

6 Evaluation

An important aspect of attribution model evaluation is recognition of the range of marketing conditions under which the evaluation is taking place. In this paper, these conditions are specified by the parameters of the DASS simulation. We define a "scenario" as the specification of a single set of DASS parameters. A "scenario family" corresponds to multiple sets of closely related parameter specifications. These specifi-

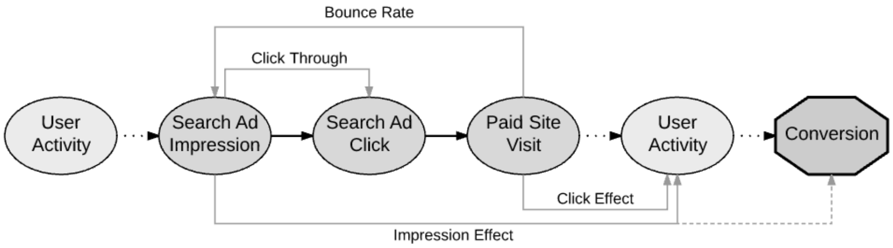


Figure 1: Diagram of the types of ad effectiveness. Serving a search ad to a user has the potential to impact user behavior in different ways. The ad may result in a direct click through to the advertiser's website, or it could have a downstream effect on the user's browsing behavior, or increase the likelihood to convert once the user is on the advertiser's website. Any of these mechanisms could lead to additional conversions.

cations differ by one, or at most two, parameter values. Most often, the parameter that is varied changes the magnitude of the ad impact so that the scenario family can be used to determine an attribution model's ability to measure ad effectiveness under different mechanisms of advertising impact, as described above.

6.1 Scenario Family Categorization

DASS is highly flexible and the list of advertising scenarios that can be created is innumerable. Although we have considered over 40 advertising scenario families, we confine the following discussion to a more limited canonical set that focuses on the search and display ad formats. These scenario families are most informative in terms of identifying primary algorithm capabilities and limitations and are sufficient for demonstrating a systematic evaluation process. They are classified into four main categories:

1. *Foundational*: Single ad channel where ad effectiveness varies.
2. *Variable Ad Effectiveness*: Ad impact decays over time and/or varies with ad frequency (i.e., there is ad burn-in and fatigue).
3. *Multiple Channel*: Situations in which multiple ad channels are present.
4. *Ad Targeting*: Ads are preferentially served to a group of users with traits or behaviors that differ from others.

6.1.1 Foundational Scenario Families

Scenario families in the foundational class include a single advertising channel with a single mode of ad effectiveness. The magnitude of ad effectiveness varies across the scenarios within this scenario family. The objective is to understand how well attribution models can capture and account for the modes of ad effectiveness described in Section 5.

1. Search Click Through: A search ad impacts user behavior by increasing the probability of a visit to the advertiser's website via a paid click. The CTR varies across scenarios in this scenario family.

2. Search with Click Effect: A search ad permanently impacts a user's downstream browsing behavior via a click on the ad. Clicking on a search ad increases the probability of performing brand related searches and visiting the advertiser's website directly. The click effect parameter varies across scenarios in this scenario family.

3. Display with Impression Effect: Exposure to (or viewing) a display ad permanently impacts a user's downstream browsing behavior by increasing the probability of searching or visiting the advertiser's website directly. The impression effect parameter varies across scenarios in this scenario family.

A foundational scenario family that we do not use in this evaluation is the scenario in which search ad impressions impact downstream user browsing behavior. Currently, search impressions are not available to attribution models. Without search impressions, no attribution model is capable of measuring the value of search when there is impression value, as Figure 6 in Sapp et al. [2016] illustrates. The scenario does not provide useful differentiating evidence across the models and so we do not include it in this model evaluation.

6.1.2 Scenario Families with Variable Ad Effectiveness

Scenario families in the Variable Ad Effectiveness category also include a single mode of ad effectiveness and a magnitude of ad effectiveness that varies across scenarios. However, in these scenario families, the impact of an ad can vary with frequency and decay over time. The objective is to understand how well attribution models work when these types of variations in ad effectiveness are present.

1. Decaying Ad Impact: A display ad impacts the downstream behavior of a user by increasing the probability of a related search or direct visit to the advertiser's website. This impact decays across time as specified by a parameter that specifies the half-life of ad impact. This half-life parameter varies across scenarios in this scenario family.

2. Burn-in: Exposure to a display ad permanently impacts downstream browsing behavior of a user by increasing the probability of searching or visiting the advertiser's website directly. However, successive ad exposures have an increasing, or decreasing, marginal impact. Burn-in is controlled by a parameter that changes the number of ad exposures required to reach the ad exposure with the maximum marginal impact on user behavior. This parameter varies across scenarios in this scenario family.

In the interest of brevity, an additional and related variable ad effectiveness scenario family is not included in this example evaluation. In this scenario family, display ads experience fatigue over multiple ad impressions. Fatigue is controlled by a parameter that changes the rate at which the marginal effectiveness of ad impressions approaches zero across multiple ad exposures.

6.1.3 Scenario Families with Multiple Channels

Scenarios with multiple channels help identify the extent to which the effectiveness of one channel might be erroneously attributed to another. In these scenarios, the ad effectiveness of one ad channel remains fixed while the other is allowed to vary.

1. Two Display Channels: Two display channels are served on different browsing states, and both channels permanently impact downstream browsing behavior of a user by increasing the probability of searching or visiting the advertiser's website directly. The magnitude of ad impact of one channel varies by increasing the impression effect parameter. Ideally, increasing the impact of this channel should not impact the conversions attributed to the second channel.

2. Two Search Channels Click Through: Generic search and branded search channels, served on different states (i.e., with independent sets of keywords), are both present in this scenario family. Both search ad channels impact user behavior by directly increasing the probability of a visit to the advertiser's website via a paid click through. The CTR of the generic search channel only varies across scenarios in this scenario family.

3. Independent Search Channels: Generic search and branded search channels, served with independent sets of keywords on different states, are both present in this scenario family. Branded search ads impact user behavior by increasing the probability of a visit to the advertiser's website via a

paid click. Generic search ads impact the downstream browsing behavior of a user by increasing branded and generic searches and increasing direct visits to the advertiser's website. Ad impact varies for the generic search channel only.

4. Search and Display Channels: Both search and display ad channels are included in this scenario family. Search ads impact user behavior by increasing the probability of a visit to the advertiser's website via a paid click. Display ads impact downstream user behavior by driving additional branded and generic searches and direct visits to the advertiser's website. Display ad effectiveness varies in this scenario family by increasing the magnitude of the impression effect parameter.

6.1.4 Scenario Families with Demographic Ad Targeting

Scenarios families with ad targeting help identify the extent to which attribution models provide useful information in the presence of ad targeting. In these scenario families, users who are more likely to be served ads are, in some way, demographically different from users who are less likely to be served ads. The magnitude of difference between these two sets of users is varied across the scenarios in each scenario family. These are challenging situations for attribution models because it is difficult to find an appropriate set of unexposed users to compare with the set of exposed users.

1. Ad Targeting For Display Ad Channel: The inherent probability of conversion is varied across two groups of users. For example, these user groups may be thought of as having different age and gender demographic distributions which result in different levels of baseline interest in the advertiser. Furthermore, the advertiser is more interested in showing ads to the user group with the higher level of baseline interest. In this scenario family, a single display ad channel can impact the downstream behavior of a user. The intensity of display ad targeting is simulated by increasing the rate at which the more targeted user group will be exposed to, and impacted by, display ads relative to the less targeted user group.

6.1.5 Other Potential Scenario Families

There are many other possible scenario families that can be considered in evaluating attribution models. We expect the number of useful scenario families to change as analysis needs grow and attribution models advance over time. However, for a scenario family to provide discriminatory value between models, at least one attribution model needs to be sufficiently capable within that scenario family. Scenario families that are not differentiating can provide insight into attribution model limitations, but are not helpful with model selection. A few examples of other potential scenario families include: the cannibalization of organic clicks by paid clicks, behavioral ad targeting (re-targeting users who have a history of interaction with the advertiser), and unobservable ad channels (e.g., offline advertising and digital channels that are beyond the attribution model's data scope). These are challenging scenarios for any attribution model.

6.2 Scoring

6.2.1 Scenario Level Scoring

Attribution ground truth is generated by running a virtual experiment, and this experiment depends on the causal measurement objective, as discussed in Section 3. In this paper, the target objective is the number of incremental conversions (IC) generated by each channel. Let x_T be the number of conversions generated by running the simulation with all ad types on, where T indicates all ads are included in the simulation. Let x_j be the number of conversions generated by running the simulation with all ad types on except for ad type b_j. Then, the IC for ad type b_j is $\delta_j = x_T - x_j$. This value is the number of conversions lost when b_j is not present.

The objective of estimating b_j is very sensible. It directly corresponds to a real world experiment that an advertiser might run to assess ad effectiveness (i.e. turn off channel b_j for a subset of users and estimate the number of conversions that are lost as a result). However, this objective does not conform to the "last event accounting principle" described in Section 2, since $\sum_j \delta_j$ usually will not match the total number of observed conversions when all channels are on. More importantly, we would like to include rules-based attribution models, which do conform to the "last event accounting principle", in our evaluations and comparisons without disadvantaging them. Consequently, we relax the objective and require attribution models to allocate the correct proportion of incremental conversions across paid ad channels.

Let x_0 be the number of conversions with all ads off. The attribution objective, relative incremental conversions for ad type j, is given by,

$$\rho_j = (x_T - x_0) \times \frac{\delta_j}{\sum_i \delta_i}. \tag{1}$$

In the evaluations below, the share of the relative IC for paid ad channels that is typically reported is given by,

$$\rho_j^{\text{share}} = \frac{\rho_j}{x_0 + \sum_i \rho_i}. \tag{2}$$

It is worth noting that ρ_j is not suitable for situations in which $\sum_i \delta_i$ is zero, or close to zero (e.g., all ad channels are completely ineffective). For these cases, we use an alternative scoring, which is defined in Appendix B.

6.2.2 Aggregate Scoring

Attribution model evaluation requires a way to measure model performance for an individual scenario and across the scenarios of a scenario family. For each paid ad channel in each scenario within a scenario family, the relative incremental conversions are computed as described above. These results are standardized before being combined. The standard error of the estimate for relative incremental conversions is found by bootstrapping the user paths. Let $\rho_{j,k}$ be the target share of incremental conversions for the j^{th} paid channel in scenario k, and let $\mathrm{SE}(\rho_{j,k})$ be the standard error of $\rho_{j,k}$. An estimate $\hat{\rho}_{j,k}$ of $\rho_{j,k}$ can be calculated for each attribution model i. The error score of model i for the j^{th} paid channel in scenario k is

$$\mathrm{e}_{j,k}^{(i)} = \frac{|\hat{\rho}_{j,k}^{(i)} - \rho_{j,k}|}{\mathrm{SE}(\rho_{j,k})}. \tag{3}$$

The scenario specific average error score for attribution model i across the J paid ad channels is

$$\mathrm{e}_k^{(i)} = \sum_j w_j \mathrm{e}_{j,k}^{(i)}, \tag{4}$$

where w_j are the ad channel specific weights that can be determined depending on the importance of each ad channel within a scenario. Therefore, for scenario family S composed of K scenarios and J paid ad channels, the error score for attribution model i is the average error across each paid ad channel and scenario,

$$\mathrm{Err}_S^{(i)} = \sum_k v_k \mathrm{e}_k^{(i)}, \tag{5}$$

where v_k are scenario specific weights. Similar to specification of w_j, the v_k can be assigned to reflect the relative importance of each scenario. These errors can be used to compare model performance for a scenario family.

In this evaluation example, we weight the J ad channels and K scenarios uniformly across scenario families. However, weighting can be non-uniformly distributed, depending on the importance of ad channels and the known, or perceived, relevance of scenarios within a scenario family. For example, it may be desirable to set w_j of Equation 4 to be proportional to the expected frequency of ad events for each channel j.

As we are primarily interested in determining which model performs best overall, scenario family errors can be further rolled up for each algorithm to rank attribution models in an evaluation (i.e. a scenario family category, or a specified group of scenario families of interest). This overall error score $Q^{(i)}$ is determined by a weighted average of the scenario family errors in an evaluation,

$$Q^{(i)} = \frac{\sum_S W_S \mathrm{Err}_S^{(i)}}{\sum_S W_S}, \tag{6}$$

where scenario family specific weights W_S can be assigned based on the known, or perceived, importance of individual scenario families. Some scenario families may be more representative of real world situations for certain advertisers, making it more appropriate to favor algorithms that perform well in those cases. For example, an advertiser whose advertising budget goes exclusively to search may be less concerned about model performance on display focused scenario families.

6.2.3 Qualitative Scoring

While the scenario level and aggregate level scoring methods are useful in ranking model performance, these metrics do not always convey the full performance story as they only measure the average error across scenarios. A more complex scoring system might be able to account for additional qualitative considerations, but for now we give them separate attention.

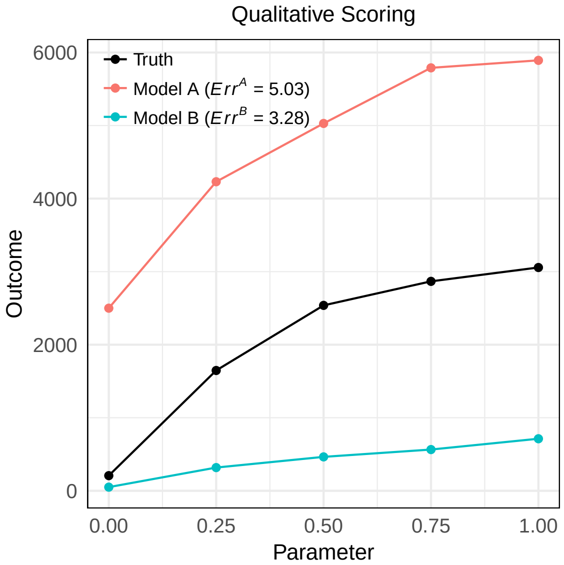


Figure 2: Illustration of the importance of qualitative considerations in assessing model performance. Although Model B has a lower scenario family error score, it does not capture the shape of the Truth curve as well as Model A.

In Figure 2, Model A captures the increasing trend of the ground truth with a relatively stable offset across all parameter values. Model B, on the other hand, does not perform well in capturing the trend, but has a lower error score. In this example, Model B outperforms Model A if only score is considered. However, Model A has desirable characteristics that are not captured in the score comparison, as it is able to appropriately track the changes in the varying parameter. Figure 2 illustrates why it is important to consider both quantitative and qualitative methods in evaluating and understanding model performance. In the scenario family examples that follow, both methods will be used in discussing model performance.

6.3 Results

We evaluate the first interaction, last interaction, linear, matched-pairs DDA (MP-DDA), and MUDDA attribution models [Help, 2017b,a, Kelly et al., 2018] with the scenario families described in Section 6.1. The first interaction model assigns full credit to the first observable event in a converting path. Similarly, the last interaction model assigns full credit to the last observable event prior to the conversion. Unlike first and last interaction, the linear model is a multi-touch attribution model in which credit is distributed evenly across all observable events preceding the conversion. The MP-DDA and MUDDA algorithms are data-driven approaches to attribution. These models compare the probabilities of conversion in converting and non-converting paths to assign credit to a target advertising event. The counterfactual gain of each event in a path (comparison of conversion probability for paths with the event compared to paths without the event) is found. The primary difference between these two models is that MP-DDA looks at all permutations of touch points for K events prior to conversion in a path [Help, 2017a], while MUDDA only considers the sequence of events upstream from the target advertising event whose impact is being evaluated. Among these models, MUDDA is the only one that is aligned with a causal view of attribution [Rubin, 1974].

For this evaluation, we assume the attribution data scope includes organic search clicks, direct navigations to the advertiser's website, display ad impressions and clicks, search ad clicks, and conversions. Search impressions are not observable. The evaluation is performed in the context of this data scope and the reporting requirements of the last event accounting principle, discussed previously, are met by all of the attribution algorithms.

Table 1 provides a quantitative summary of model performance across scenario families. Algorithm errors $\mathrm{Err}_S^{(i)}$ are reported for each scenario family and the overall error score $Q^{(i)}$ is determined using equal weighting across scenario families. The causal attribution model, MUDDA, performs best in most scenarios.

As noted previously, it is important to look beyond the evaluation scores to better understand the relative and absolute model performance. So, next we include a qualitative discussion of performance for each scenario family.

Foundational Scenarios. In the first two scenario families, only the search ad channel is present and we vary the level of ad effectiveness in different ways. In the first scenario family, we vary click through rate. No impression or persistent click effects are present. The CTR parameter controls the rate that users click through to the advertiser's website from a search ad. This scenario family evaluates model performance when an ad has a direct and immediate impact on user behavior with no subsequent downstream impact. In the second scenario, we vary the impact of an ad click on downstream browsing behavior. There is a small fixed CTR and no impression effect. In this scenario family, clicks on search ads increase the likelihood of brand related searches and visits to the advertiser's website by organic clicks and direct navigations.

Figure 3 is a plot of the true incremental conversions and each algorithm's attributed conversions for different search click through rates. As the ads become more effective, the number of true incremental conversions increases. All of the attribution models do relatively well in tracking the true number of incremental conversions over all levels of CTR. Little differentiation is seen between the models in this scenario family since there is only one paid channel for the models to attribute credit. On its own, this scenario family is not particularly useful for differentiating model performance for this evaluation, perhaps with the exception of the MP-DDA model, which has a larger discrepancy compared to other models. However, it is worthwhile to consider this scenario because it is a simple canonical situation that an attribution model is expected to handle well. The opportunity for more differentiation across models exists when there is more than one search channel. This is illustrated below in the multiple channel scenario family category.

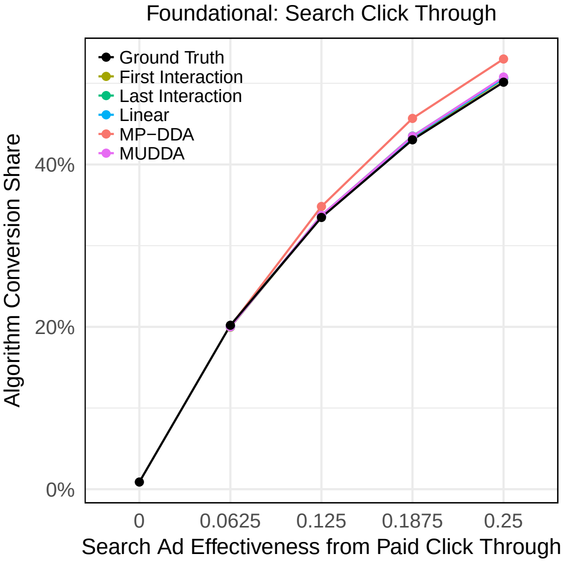


Figure 3: Foundational scenario family with one search ad channel. This plot shows the attributed IC share to search as search ad click through rate varies across scenarios.

The second search channel scenario, illustrated in Figure 4, indicates that most models are unable to recognize the change in incremental conversions for increasing levels of ad effectiveness. First interaction, linear, and MUDDA are able to slightly capture the upward slope, but MP-DDA and last interaction perform poorly. MP-DDA has a large positive offset but tracks the changes in varying ad effectiveness. Last interaction is unable to capture the trend across scenarios, as this model is incompatible with a

	First Interaction	Last Interaction	Linear	MP-DDA	MUDDA
Search Click Through	0.84	1.07	0.94	4.01	1.1
Search with Click Effect	0.79	2.98	1.47	7.68	1.23
Display with Impression Effect	13.98	7.54	4.54	5.97	1.05
Decaying Display Ad Impact	15.05	2.55	7.57	2.31	1.22
Display Burn-in	20.8	2.35	10.92	2.16	1.5
Two Display Channels	11.87	6.09	3.83	4.69	0.59
Two Search Channels Click Through	2.24	0.76	1.43	1.61	1.43
Independent Search Channels	0.84	2.2	1.05	4.75	0.79
Search and Display Channels	13.26	5.23	5.22	4.42	0.92
Display Ad Targeting	7.93	3	3.29	1.61	0.92
Overall Error	8.76	3.38	4.03	3.92	1.07

Table 1: Scenario family errors and overall error by algorithm. Cells highlighted in gray indicate the algorithm with the lowest error score $Q^{(i)}$. Overall, MUDDA performs best in this evaluation.

mechanism of ad effectiveness that affects downstream browsing behavior. The performance of MUDDA can be explained by a combination of missing information due to scoping, reporting restrictions and "user browser dissimilarity", as introduced in Section 4.2 of Sapp and Vaver [2016]. These are all areas to consider in the process of improving and better understanding model performance.

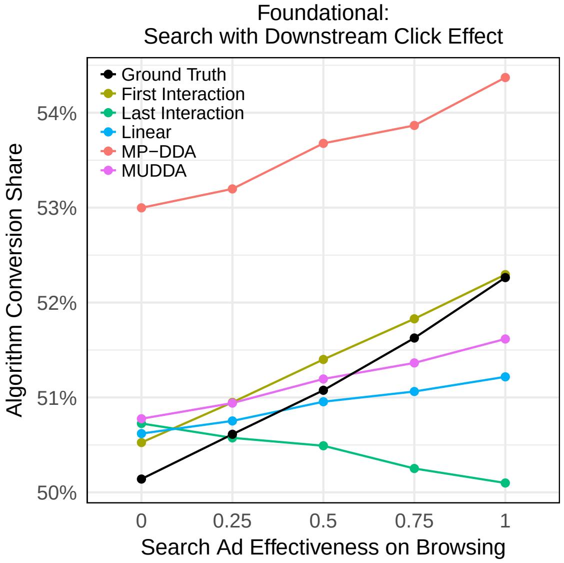


Figure 4: Foundational scenario family with one search ad channel that has downstream impact on browsing behavior. This plot shows the attributed IC share to search as persistent ad effectiveness from a search click through varies across scenarios.

Next, we consider the case in which display ads impact the downstream browsing behavior of users. After a display ad impression, users may become more likely to perform branded and generic searches and visit the advertiser's website through direct navigation.

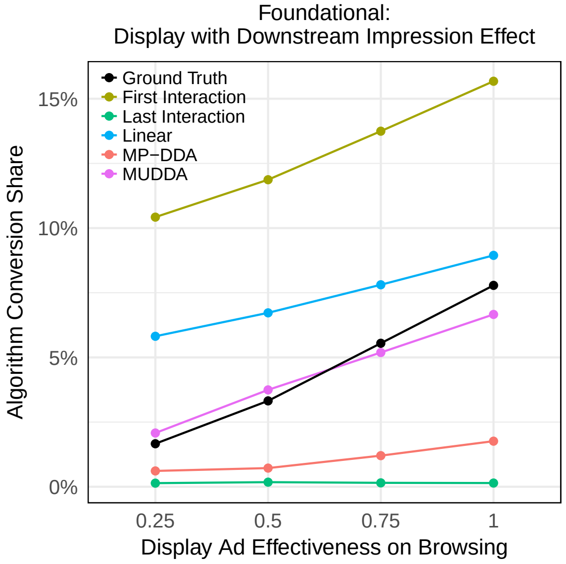


Figure 5: Foundational scenario family with one display ad channel that has downstream impact on browsing behavior. This plot shows the attributed IC share to display as ad effectiveness from display impressions vary across scenarios.

Results from this scenario family are captured in Figure 5. Last interaction is unable to capture the incremental conversions from the display ad because an ad impression can never be the last interaction prior to a conversion. There will always be a site visit between an impression and a conversion and the model attributes full credit to the event immediately before the conversion. First interaction is able to capture the general trend of the incremental conversions, but has a very significant offset. This model recognizes the role of attribution, but it is not capable of identifying the extent to which the impact generates incremental conversions. The linear model performs better than the other rules-based models, but it still does not quite capture the IC share. The MP-DDA model performs very poorly, which is a result of the downstream matching that the model performs. In this scenario, ads have downstream impact on browsing behavior, yet MP-DDA requires the downstream paths to be the same. Only the MUDDA model is able to capture the causal impact of downstream ad effectiveness reasonably well, as it matches on the paths upstream of the ad event only.

When a model doesn't perform well in a foundational scenario family it is an indication that the model will also have trouble in more complex scenarios. We don't expect the addition of complexity to fix the more fundamental shortcomings of a model.

Variable Ad Effectiveness Scenarios. We present results for two scenario families that vary the rate at which ad impact decays across time and saturates with ad frequency. These scenarios have a single display ad channel in which advertising has a downstream impact on user browsing behavior and a small click through rate. After an impression, ad impact can result in users being more likely to perform generic and brand related searches and more likely to visit the advertiser's website through organic browsing activity.

In the first scenario family, the rate at which display ad impressions lose effectiveness across time is varied. As the half-life of ad impact increases, ads have a more sustained impact on user behavior and ads generate more incremental conversions, as shown in Figure 6.

Qualitatively, the model performance is similar to the results of the foundational display ad scenario family described previously. MUDDA performs reasonably well, while the other models do not. Yet, in Table 1, MP-DDA and last interaction are ranked lower than first interaction and linear. This is a result of larger vertical offsets for the linear and first interaction models, rather than a true change in the capabilities of these models to respond to changes in ad effectiveness. This scenario family demonstrates the importance of considering both quantitative and qualitative performance in understanding the capabilities of a model.

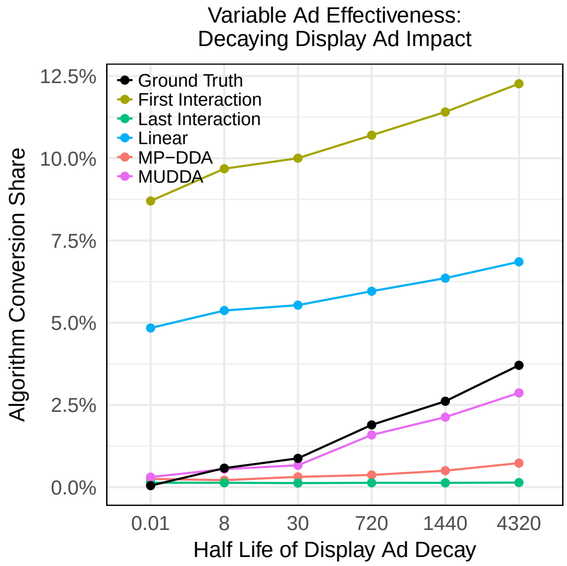


Figure 6: Variable ad effectiveness scenario family with a single display ad channel that impacts downstream browsing behavior. This plot shows the attributed IC share as the half-life of display impressions increases.

In the second variable ad effectiveness scenario family, we consider the impact of varying the burn-in of display impressions on model performance. A parameter is varied that controls the number of ad exposures required to reach maximum marginal impact. Similar to the previous scenario family, display ads have a small CTR and change downstream browsing behavior.

Figure 7 illustrates that only MUDDA is somewhat capable of recognizing ad effectiveness that diminishes with frequency. Although MUDDA does not account for this diminished ad effective-

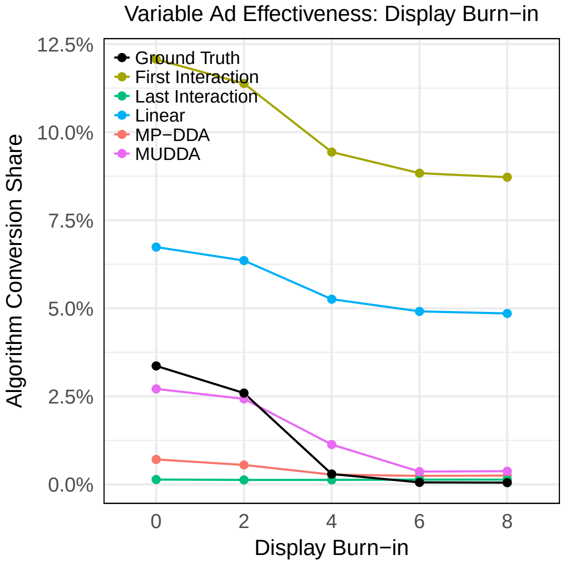


Figure 7: Variable ad effectiveness scenario family with a display ad channel that impacts downstream browsing behavior. This plot shows the attributed IC share as the burn-in of display ad impressions varies across scenarios. The x-axis indicates the number of ad impressions required to reach the maximum marginal effect on user behavior.

ness at the event level, its matching mechanism inherently captures the impact.

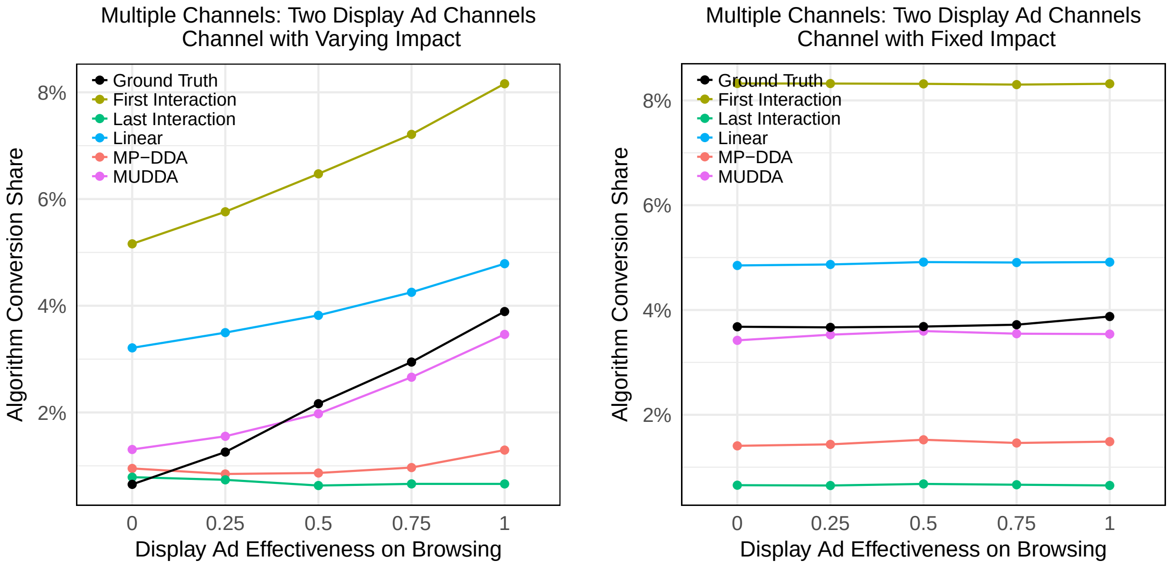
Multiple Channel Scenarios. We consider four scenario families to assess model performance in the presence of multiple channels. In the first scenario family, two display ad channels are served on different user activity states. The effect of an impression in one channel is held constant while the impression effect in the second channel is varied. Since the campaigns are served on different activity states, the channels operate independently. We expect similar results to the Foundational scenario family with one display ad (see Figure 5), but the primary objective is to determine if the increasing effectiveness of one channel is misattributed to the channel with fixed ad effectiveness.

The results from this scenario are illustrated in Figure 8. Figure 8(b) shows that the display channel with fixed ad impact does not take IC share credit for the display channel with variable ad impact. With respect to model ranking, these results closely resemble the foundational case in which only one display channel is present. Again, MUDDA is the only algorithm capable of capturing the downstream impact of the display channel with varying ad effectiveness when two channels are present in the simulation.

Since the ads are served on different activity states, the channels do not interact with each other and Figure 8(b) does not provide additional information about algorithm performance. The models either consistently overestimate (first interaction and linear), underestimate (MP-DDA and last interaction), or accurately estimate (MUDDA) the IC share for both channels across all levels of ad effectiveness. However, in a scenario family in which multiple ad channels are served on the same state, we expect the channels to interact with each other and the models to misattribute credit. We do not include this scenario family in the evaluation as the models do not have enough information to perform well in this situation and so the results will be non-differentiating across models.

Next, we consider the case in which two search channels are present; search ads placed against branded search terms and search ads placed against generic search terms. Similar to the previous scenario family, these channels are independent and served on unique activity states so that, effectively, these search campaigns have different sets of keywords. This scenario family is analogous to the Foundational Search Click Through scenario family. In this scenario family, the click through rate is fixed for the branded search channel and the click through rate varies across scenarios for the generic search channel.

Figure 9 shows that all models are able to perform well in attributing credit to the generic search channel. Table 1 indicates that there is more differentiation between models compared to the search CTR scenario with one channel only, and that last interaction model performs best. This scenario family is an example of an advertising situation in which last interaction may outperform the other models, as the mechanism of ad effectiveness is a direct click through to the advertiser's website. The branded



(a) This plot shows the attributed IC share to the display ad channel that has varying ad impact from an impression.

(b) This plot shows the attributed IC share to the display ad channel that has fixed ad impact.

Figure 8: Multiple channel scenario family with two display ad channels that have downstream impact on browsing behavior. The plots indicate that the effectiveness of one channel is not misattributed to the second display channel with fixed ad effectiveness.

search channel does not take credit away from the generic search channel with varying click through rate and so Figure 9 is sufficient to illustrate attribution model performance.

The third scenario family with multiple channels also has two independent search channels; search ads placed against branded search terms and search ads placed against generic search terms. In this case, the generic search ad has a fixed click through rate and we vary the magnitude of impact that an ad click has on downstream user browsing behavior. The branded search channel has a fixed click through rate and no downstream impact on user browsing behavior. As in the scenario family with two display channels, the primary objective is to determine if the increasing effectiveness of one channel is misattributed to the channel with fixed ad effectiveness.

Figure 10 indicates that most of the models are able to recognize the increasing downstream impact from generic search ads. As in the single search channel scenario family, the last interaction algorithm has the most trouble capturing the upward trend of the incremental conversions. Since the ads are served independently of each other, we observe again that the branded search channel with fixed ad impact does not take credit for the ad impact of the generic search channel with varying effectiveness (plot is not shown).

This scenario family can also be developed with the search channels served on the same state, and therefore, “competing” with each other. This would be the case if the campaigns use overlapping sets of keywords. However, this scenario family is not included in the evaluation as the attribution models do not have information to understand this overlap. More specifically, if one channel is turned off there may be no impact to the number of overall conversions. This may be due to the ineffectiveness of this channel, or due to an overlapping channel with similar ad effectiveness showing ads in place of this channel. With the modeling objective described in Section 3, there is no way for an attribution model to recognize the difference.

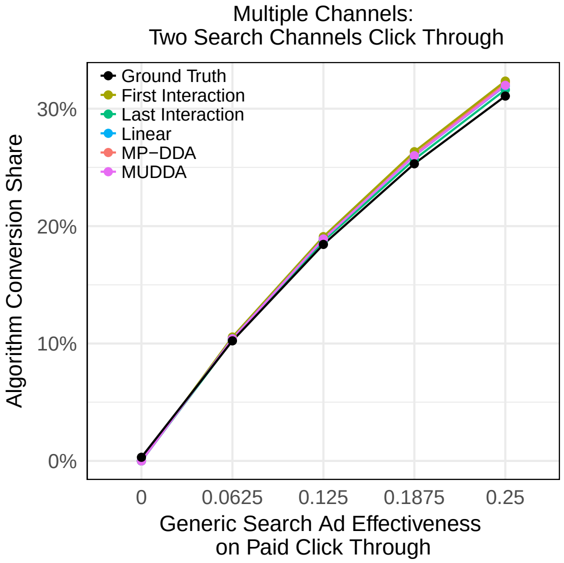


Figure 9: Multiple channel scenario family with independent "branded" and "generic" search channels that do not have any downstream ad effectiveness. This plot shows the attributed IC share to the generic search channel as the generic search click through rate varies across scenarios.

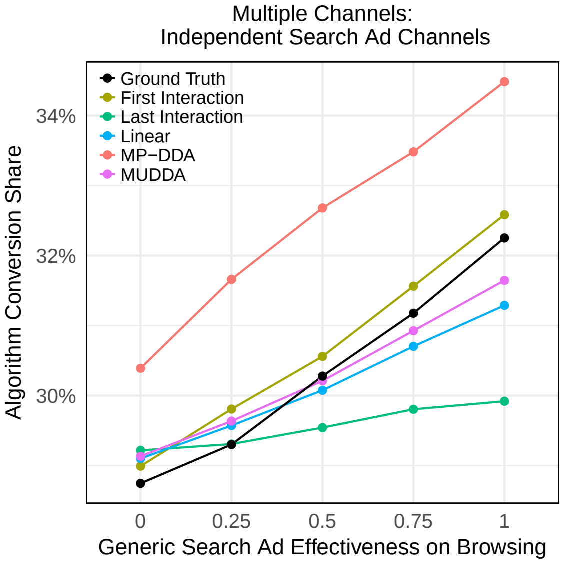


Figure 10: Multiple channel scenario family with independent "branded" and "generic" search channels. This plot shows the attributed IC share to the generic search channel, which has a downstream impact on browsing behavior through search click throughs that is varied along the x-axis.

The final multiple channel scenario family we consider has a display and search channel. In this scenario family, the search channel has a fixed CTR and no downstream ad impact. Display impressions can increase the user's propensity to conduct generic and branded searches and visit the advertiser's website from organic activity. The impact from these impressions changes across the scenarios of this family.

The results of this scenario family, shown in Figure 11, closely follow the results from the foundational display scenario family. As observed in the previous two scenario families with two ad channels, the models are able to discriminate between the effectiveness of channels that are served on different activity states. In this case, the search ad channel does not erroneously take credit for display channel effectiveness. Similarly, in an additional scenario family not included in this evaluation, we observe that the display ad channel does not take credit from a search ad channel when search ad impact varies and display impact is fixed.

Demographic Ad Targeting Scenarios. The last scenario family we present illustrates the impact of ad targeting when display ads are preferentially served to one group of users over another. In this scenario family, there are two groups of users. The user groups may have different baseline propensities to convert and propensities to be served and impacted by ads. The degree of ad targeting is controlled by varying these differences between the user groups, which users receive ads, which users are impacted by ads, and which users are more likely to convert. These differences are varied to change the degree of ad targeting.

Figure 12 indicates that most models are unable to capture the user heterogeneity in the data. First interaction, linear, and MUDDA appear to track the increase in incremental conversions, but since these models do not take into consideration heterogeneous user sets, the models exhibit an offset from the true IC share and it is unlikely they will perform well in other ad

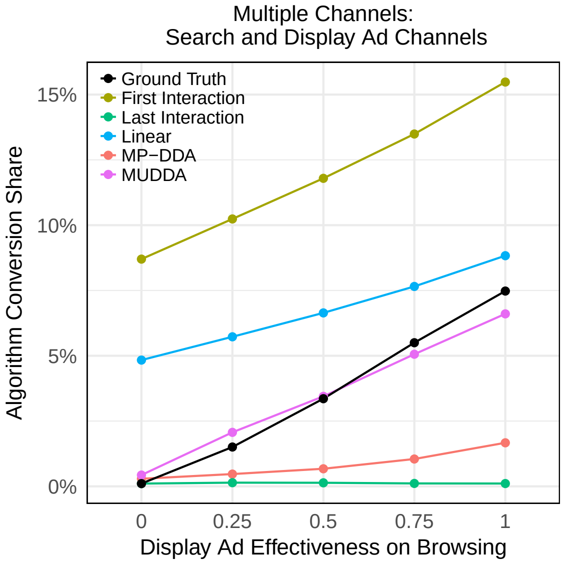


Figure 11: Multiple channel scenario family with one display and one search ad channel. The search ad channel has a fixed click through rate and the display ad channel downstream browsing behavior. This plot shows the attributed IC share to display as display ad effectiveness varies across scenarios.

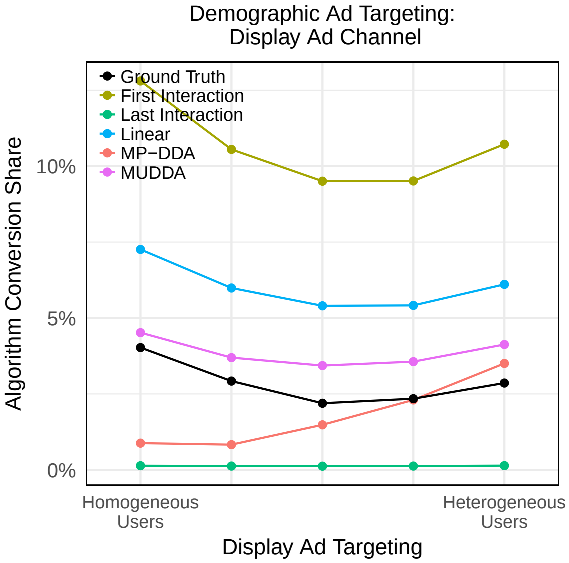


Figure 12: Demographic ad targeting scenario family with one display ad channel that has a downstream impression impact on user browsing behavior. The two user groups become more distinct moving from left to right. This plot shows the attributed IC share to display as the magnitude of ad targeting increases.

targeting situations. MUDDA performance deteriorates as ad targeting increases because, in matching the exposed and unexposed groups, the model can not take into consideration the inherently different conversion rates between users. As there is insufficient information in the attribution data, this scenario may not be worthy of inclusion in this evaluation.

7 Conclusion

In order to evolve and remain relevant, attribution modeling needs an impartial model evaluation process with a clearly defined causal measurement objective. This process includes quantifying a model's ability to handle canonical advertising scenarios and a means for aggregating results across these scenarios. It also includes a hierarchical structure for organizing scenarios that helps to put results into perspective and facilitate algorithm improvement.

In this paper, we outline an evaluation process based on the Digital Advertising System Simulator (DASS). DASS is an ideal tool for generating path data for evaluation scenarios due to its flexibility and its ability to generate the truth needed for evaluation. The example set of scenarios that are described here are by no means complete. Any catalog of useful scenarios will undoubtedly grow and evolve over time. New capabilities will be added to the simulator, attribution models will improve, new data sources will be made available to attribution models, and new questions about model capabilities will be posed.

Attribution model evaluation should be viewed as an emerging process. For example, results from a few of the scenario families presented in this paper demonstrated model sensitivity to parameter settings and scoping. The model evaluation process can be improved by a sensitivity analysis that considers a wider range of DASS simulation parameters and degrees of missing data.

Through the scenario families presented above, we see evidence that models adhering to a causal view of attribution are likely to perform best in estimating the number of incremental conversions generated by marketing activity. However, a causally based model is no assurance of performance. These models are still limited by the completeness of data sources, reporting requirements, and campaign implementation. There is plenty of room for improvement in attribution modeling. A systematic evaluation process is the best way to determine the strengths and challenges that attribution modeling must overcome to continue to be a useful and trusted source of measurement.

Acknowledgements

The authors would like to thank many people at Google for their guidance and support. Many thanks to Tony Fagan, David Chan, Jim Koehler, and Georg Goerg for their helpful feedback. We would also like to thank Penny Chu and members of the engineering and product teams that work on digital attribution for many related discussions and contributions, support, and their commitment to a systematic evaluation process.

Appendix A Time-Based Simulator

This appendix describes an updated version of DASS that takes into account the inter-arrival time of user activities and generalizes the approach for specifying ad impact across multiple ad exposures. Scenario families from the Variable Ad Effectiveness category presented in Section 6.1.2 were developed with this version of the simulator.

A.1 Scale Function

In the initial implementation of the DASS simulator, user behavior in the absence of advertising is specified with a baseline transition matrix. An ad serving event may impact downstream user behavior through a scale factor that inflates one, or more, transition probabilities of the baseline matrix. For the time-based simulator the scaling factor evolves across ad frequency and time, as described in Appendix A.3. A classical S-curve function is used to model user response to advertising. The scale factor, S_k, is dependent on the number of ad events shown to the user, n_k, up to the current point in time for channel k. After the scale factor is applied, the transition matrix is renormalized and the scaled transition matrix is used to determine the next user activity.

The scale factor generated by channel k with n_k ads served is given by,

$$S_k(n_k) = a(-1 + \frac{2}{1 + \exp^{-b(n_k - n_0)}}) + 1 + c \quad (7)$$

where n_0 is the number of ad serving events that has the largest impact on the transition probabilities and a, b, and c are additional parameters that specify the S-curve function. The following specifications are used to find a, b, and c:

1. $S_k(0) = 1$.
   Before any advertising events from channel k, user behavior is governed by the baseline transition matrix.

2. $S_k(\infty) = S_{\max_k}$.
   The largest scale factor that can be generated by channel k is bounded by a maximum value of $S_{\max_k}$.

3. $R_{\max_k} = \frac{dS_k}{dn_k} S_k(n_0)$.
   The maximum scale factor change per ad served for channel k occurs at n_0. Prior to reaching n_0, increases in n_k generate an increased scale factor change per ad. After reaching n_0, increases in n_k generate a diminished scale factor change per ad. The maximum change is $R_{\max_k}$.

The specification of n_0, $S_{\max_k}$, and $R_{\max_k}$ defines the scale function. In the context of variable ad effectiveness, n_0 determines the number of ad exposures required to reach the maximum marginal ad impact (ad burn-in), and $R_{\max_k}$ controls the rate at which marginal ad impact diminishes (ad fatigue).

For example, in the display burn-in scenario family presented in Figure 7, n_0 varies and $S_{\max_k}$ and $R_{\max_k}$ are fixed. For the case in which $n_0 = 0$, $S_{\max_k} = 3.375$, and $R_{\max_k} = 0.5$, the corresponding values of a, b, and c are 2.375, 0.421, and 0, respectively. In this scenario, the scale functions for various values of n_0 are shown in Figure 13.

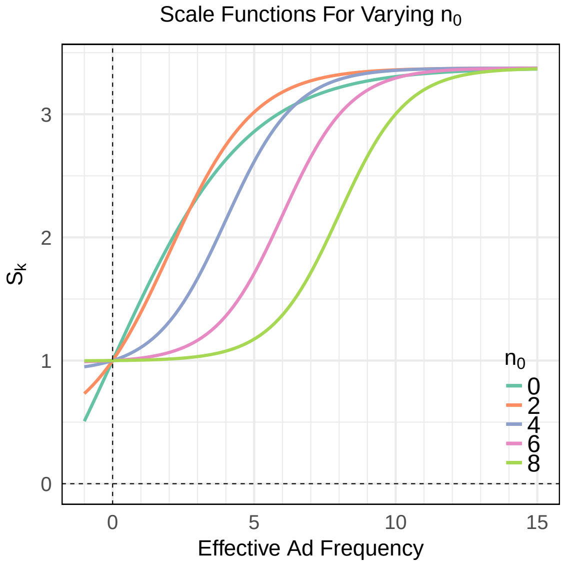


Figure 13: Illustration of the scale functions for n_0 set to $0, 2, 4, 6$, and 8 in the display burn-in scenario family. As n_0 gets larger, marginal ad impact reaches $R_{\max_k}$ at a slower rate.

Figure 13 indicates how components of the transition matrix are scaled as a function of the ad frequency. The curves are bounded by the maximum scale factor 3.375. For each curve, the greatest scale change per ad event, 0.5, is reached at the n_0^{th} ad.

A.2 Time Dependence

Allowing the impact of an ad event to vary over time is accomplished by defining a set of parameters that govern the inter-arrival time between events, and by tracking the "effective frequency" of ad exposures. Before describing these parameters, we explain how time is incorporated into the simulation.

Over time, users engage and unengage with a category or brand, as indicated by their browsing activities. For example, searching or visiting an advertiser's website is a more engaged activity than visiting a third-party website. In the simulation, user activity states are characterized as either "engaged" or "unengaged", and we expect the time between activities to vary based on these labels. This characterization is used to limit the number of parameters needed to control the inter-arrival times in the simulation, which is especially important when a simulation includes a large number of states. Additionally, the time between a search and a click-through to the advertiser's website is modeled separately, since this transition is expected to occur on a shorter time scale.

For each type of transition, (i.e., engaged activity to engaged activity, engaged to unengaged, unengaged to engaged, and unengaged to unengaged), the inter-arrival time $\tau = \log_{10}(t_{i+1} - t_i)$ is sampled from a linear combination of D Gaussian distributions that depend on the type of transition. For example, for a transition from an unengaged activity state to another unengaged activity state, τ is sampled from the following distribution,

$$p_{uu}(\tau) = \sum_{d=1}^{D} \alpha_{uu_d} \mathcal{N}(m_{uu_d}, \sigma^2_{uu_d}) \tag{8}$$

where subscript uu indicates an unengaged to unengaged state transition, $\sum_{d=1}^{D} \alpha_{uu_d} = 1$, and each component d may have a unique mean and variance specification. Unique distributions that follow the form of Equation 8 are specified for the following transitions:

- Unengaged to unengaged state (p_{uu})
- Unengaged to engaged state (p_{ue})
- Engaged to unengaged state (p_{eu})
- Engaged to engaged state (p_{ee})
- Search to site visit state, or search click-through, (p_c)

Each distribution may have a unique set of means, variances, and mixture probabilities. In the simulation, there is a joint dependence of inter-arrival time and the type of consecutive activity pair, which requires the inter-arrival time and the second activity in the pair to be generated simultaneously. This process is described below.

A.3 Determining Ad Impact Over Time

Ad impact is modelled to allow the transition matrix scaling factor to increase with ad exposure and decrease with lack of ad exposure over time. This is accomplished using an "effective frequency" of ad exposure, $\hat{n}$, which is tracked across each user's path for each paid ad channel. This parameter is dependent on the previous activities in the user path, the elapsed time between ad events Δt, and the half-life of the effective frequency of ad exposures, t_d. The parameter t_d specifies the rate at which the impact of ad exposures diminish across time.

Prior to the start of the simulation, the scale function for channel k is fixed and determined by the specified values for n_0, $R_{\max_k}$, and $S_{\max_k}$. Values are also specified for the means, variances, and mixture probabilities of the set of inter-arrival time distributions, and the half-life of the effective frequency. Additionally, the activity states $a_1, \ldots, a_n$ are categorized as "engaged" or "unengaged". The updated simulation that includes time dependent ad impact is described below for a single user stream.

0. At the start of the simulation, prior to any ad serving events, user activity is determined by the baseline transition matrix. The effective frequency of ad exposures is set to zero ($\hat{n}_{previous} = 0$). It follows that $S_k(\hat{n} = 0) = 1$, indicating no change to the baseline transition matrix.

1. At the first ad serving event, set $\hat{n} = 1$.

2. Determine the next activity and sample the corresponding inter-arrival time between activities, t_a, jointly through the steps outlined below.

   (a) Determine whether the current activity state is categorized as "engaged" or "unengaged".

   (b) Accordingly, sample from the two relevant distributions described in Section A.2 to find the two possible inter-arrival times. One time, t_e, assumes a transition to an engaged activity and the other time, t_u, assumes a transition to an unengaged activity.

   (c) For each inter-arrival time, t_e and t_u, compute the effective frequency and the associated scaling factor in order to find the updated transition matrices, M_e and M_u, respectively. The effective frequency is found by,

$$\begin{aligned} \hat{n}_{updated} &= f(\Delta t, \hat{n}_{previous}) \\ &= \hat{n}_{previous}(1/2)^{\Delta t/t_d}, \end{aligned} \tag{9}$$

   where $\Delta t = t_a$ is the time since the first ad was shown and t_d is the half-life of the effective frequency specified at the start of the simulation. Using the corresponding transition matrices, compute the probability of transitioning to an engaged activity, p_e, and the probability of transitioning to an unengaged activity, p_u. These probabilities are found by summing the transition probabilities of the unengaged or engaged states corresponding to the row of the current activity state.

   (d) Use the normalized engaged or unengaged probabilities, $p_e/(p_e + p_u)$ or $p_u/(p_e + p_u)$, to determine whether the second activity type, $\bar{a}$, will be sampled from the "engaged" or "unengaged" set of states. For example, to determine if the second activity type is "engaged", $Bernoulli(\frac{p_e}{p_e+p_u}) \equiv 1$. This sampling also determines which of the two inter-arrival times and updated transition matrices to use.

   (e) Find the set of transition probabilities, $q_1, \ldots, q_n$, that correspond to a transition to an activity of type $\bar{a}$ with the updated transition matrix.

(f) Determine the second activity by sampling from $Multinomial(1, \pi)$, with event probabilities $\pi = (q_1 / \sum_i q_i, \ldots, q_n / \sum_i q_i)$.

3. Update the effective frequency with Equation 9.

4. Scale the relevant transition probabilities in the baseline transition matrix by $S_k(\hat{n}_{updated})$ and renormalize in order to determine the next activity in the user path. If multiple media channels are included in the simulation, scaling of the transition matrix entries will be the product of the scalings from each media channel. More detail is provided in Section A.4.

5. Find the next activity, $\bar{a}$, and the corresponding inter-arrival time, t_a, between the previous user activity and the upcoming user activity according to Step 2. If an ad is served, set $\hat{n}_{previous} = \hat{n}_{updated}$ and set $\hat{n}_{updated} = \hat{n}_{previous} + 1$.

6. Find Δt, the time since the last ad was served. Since there may be multiple activities between ads, $\Delta t \geq t_a$ always holds.

7. Update the effective frequency for the ad exposure using Equation 9, where Δt is now the time since the last ad was served.

8. Repeat Steps 4-7 until the absorbing activity state (typically "end of session") is reached.

A.4 Compounding Ad Impact From Multiple Channels

When multiple ad channels are included in the simulation, the impact from the scale factors can be combined to create a compounded effect on the baseline transition matrix. The compounding impact of cross-channel advertising is controlled by the following procedure:

1. Find the scaling factor S_k for each channel.

2. Let S be the largest transition matrix scaling across all ad channels, and let P be the product of these channel level scaling factors.

3. Specify a fixed (global) parameter, $\beta \in [0, 1]$, that is used to control the compounding effects of ads from multiple channels. When $\beta = 0$, there is no compounding effect, and when $\beta = 1$, there is a complete compounding effect of ads from multiple channels.

4. Specify a fixed (global) parameter, S^*, which can further limit the impact of ads across all channels.

5. Use the scaling factor $s = \min[S + \beta \times (P - S), S^*]$ to scale the baseline transition matrix (as in the single channel case), and renormalize.

Appendix B Relative Incremental Conversions

The target attribution objective, incremental conversions (IC), was defined in Section 6.2.1. Due to reporting restrictions, no channel can be assigned a negative credit, even an ineffective one. So, when at least one $\delta_j < 0$, the formula for the relative share of incremental conversions ρ_j^{share} given in Equation 2 must be modified. This appendix describes IC scoring for situations in which the absolute IC is less than zero for one or more ad channels in the simulation.

Classify the absolute incremental conversions for each ad type b_j into one of three disjoint sets:

$$S^0 = \{x_j | \delta_j = x_T - x_j = 0\}$$
$$S^- = \{x_j | \delta_j = x_T - x_j < 0\}$$
$$S^+ = \{x_j | \delta_j = x_T - x_j > 0\}$$

In order to compute a relative IC, $S^+ \neq \emptyset$[1].

[1] The situation in which $S^+ \neq \emptyset$ can happen in practice, and therefore these scenarios are worthy of consideration. For evaluations that include these scenarios, it is best to use a scoring metric based on the absolute number of IC generated by each channel that does not attempt to comply with reporting requirements that preclude negative credit.

When $S^+ \neq \emptyset$ and $S^- = \emptyset$, the original formula in Equation 2 can be used to find the relative share of IC for each b_j. For the case in which $S^+ \neq \emptyset$ and $S^- \neq \emptyset$, the relative IC share are computed by rescaling the absolute IC in S^+ and S^-, as described below. Let f_j be the target share of incremental conversions, similar to the ρ_j^{share} from Equation 2.

The share of relative IC that occur in the absence of observed ads in the simulation is $f_0 = x_0/x_T$. To find the relative IC share for each observed ad channel, define

$$\Delta^- = \sum_{i:\delta_i<0} \delta_i$$

$$\Delta^+ = \sum_{i:\delta_i>0} \delta_i$$

To account for the negative contributions from S^-, the relative IC share for ads in S^+ are inflated by the average size of δ_i in S^-, which is $\bar{\delta}^- = \frac{|\Delta^-|}{|S^-|}$. Then, the true IC share for each paid ad channel is

$$f_j = \begin{cases} 0 & \{\delta_j \in S^0\} \\ (1 - f_0) \times \frac{(\delta_{\text{ALL}} + \bar{\delta}^-)}{\delta_{\text{ALL}}} \times \frac{\delta_j}{\Delta^+} & \{\delta_j \in S^+\} \\ (1 - f_0) \times \frac{\bar{\delta}^-}{\delta_{\text{ALL}}} \times \frac{\delta_j}{|\Delta^-|} & \{\delta_j \in S^-\} \end{cases}$$

where $\delta_{\text{ALL}} = x_T - x_0$ is the total IC from all ads. With this definition, $f_0 + \sum f_j = 1.0$, since $\sum_{\delta_j \in S^+} \frac{\delta_j}{\Delta^+} = \sum_{\delta_j \in S^-} \frac{\delta_j}{\Delta^-} = 1$.

B.1 Example

This section illustrates a calculation of the relative incremental conversions when the IC of at least one ad channel in the simulation is not positive. Suppose we have four ad channels present in the simulation with the following absolute number of conversions associated with each ad type,

	x_0	x_1	x_2	x_3	x_4	x_T
Conversions	50	90	80	105	110	100

Table 2: Absolute conversions in virtual experiments.

In this simulation, the disjoint sets of ad types are: $S^0 = \emptyset$, $\{x_3, x_4\} \in S^-$, and $\{x_1, x_2\} \in S^+$. Therefore,

$$\begin{aligned} \Delta^- &= (x_T - x_3) + (x_T - x_4) = -15 \\ \Delta^+ &= (x_T - x_1) + (x_T - x_2) = 30 \\ \bar{\delta}^- &= |-15|/2 = 7.5 \\ \delta_{\text{ALL}} &= x_T - x_0 = 50 \end{aligned}$$

The relative IC share for non-ads is $f_0 = 50/100 = 0.5$ and the relative IC share for ad types 1, 2, 3, and 4 are:

$$\begin{aligned} f_1 &= 0.5 \times \frac{(50 + 7.5)}{50} \times \frac{10}{30} = 0.1917 \\ f_2 &= 0.5 \times \frac{(50 + 7.5)}{50} \times \frac{20}{30} = 0.3833 \\ f_3 &= 0.5 \times \frac{(-7.5)}{50} \times \frac{5}{15} = -0.025 \\ f_4 &= 0.5 \times \frac{(-7.5)}{50} \times \frac{10}{15} = -0.05 \end{aligned}$$

References

David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. Evaluating online ad campaigns in a pipeline: Causal models at scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 7–16, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. doi: 10.1145/1835804.1835809. URL http://doi.acm.org/10.1145/1835804.1835809.

Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, ADKDD '12, pages 7:1–7:9, 2012. ISBN 978-1-4503-1545-6. doi: 10.1145/2351356.2351363. URL http://doi.acm.org/10.1145/2351356.2351363.

Analytics Help. Data-driven attribution methodology, 2017a. URL https://support.google.com/analytics/answer/3191594?hl=en.

Google Analytics Help. About the default attribution models, 2017b. URL https://support.google.com/analytics/answer/1665189?hl=en&ref_topic=3205717.

Joseph Kelly, Jon Vaver, and Jim Koehler. A causal framework for digital attribution. Technical report, Google LLC, 2018. URL https://ai.google/research/pubs/pub46905.

Alice Li and P. K. Kannan. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. 51:40–56, 02 2014.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66 (5):688, 1974.

Stephanie Sapp and Jon Vaver. Toward improving digital attribution model accuracy. Technical report, Google Inc., 2016. URL https://research.google.com/pubs/pub45766.html.

Stephanie Sapp, Jon Vaver, Minghui Shi, and Neil Bathia. Dass: Digital advertising system simulation. Technical report, Google Inc., 2016. URL https://research.google.com/pubs/pub45331.html.

Xuhui Shao and Lexin Li. Data-driven multitouch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 258–264, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020453. URL http://doi.acm.org/10.1145/2020408.2020453.