A Causal Framework for Digital Attribution

Joseph Kelly, Jon Vaver, Jim Koehler

Google LLC

Abstract

In this paper we tackle the marketing problem of assigning credit for a successful outcome to events that occur prior to the success, otherwise known as the *attribution problem*. In the world of digital advertising, attribution is widely used to formulate and evaluate marketing but often without a clear specification of the measurement objective and the decision-making needs. We formalize the problem of attribution under a causal framework, note its shortcomings, and suggest an attribution algorithm that is evaluated via simulation.

1 Introduction

Advertisers have a primary need to know how best to allocate their advertising resources to maximize the return on their investment. In digital advertising, this often translates into estimating how effective advertising campaigns are at increasing the number of purchases made by consumers. This measurement is frequently estimated using digital attribution. (Analytics, 2018)

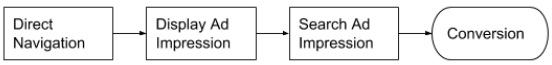

Figure 1: Consumer Path Example

Digital attribution is the process of assigning credit for a successful outcome (known as a conversion) to observed digital consumer engagements that occurred prior to the converting event. For example, in Figure 1, the advertiser would like to know how much credit for the conversion should be shared between the *Display Ad Impression*, *Search Ad Impression* and a *Direct Navigation* to the website. This event level attribution credit may be aggregated across paths or other event level features, such as a keyword or geographic location, and used by advertisers to make decisions about allocating funds.

Advertisers have many choices when it comes to attribution modeling. The simplest are the rules-based models: the last event model, which gives all credit to the event just prior to the conversion, the first event model, which give all credit to the first observed event, and the linear model, which assigns equal credit across all observed events (Analytics, 2017). There are other more complex models, including various forms of *data driven attribution*, such as the one we describe in this paper.

2 Issues With Current Attribution Practices

Although popular, and easy to use in practice, there are significant drawbacks related to the rule-based models and current attribution practices. The first is that the concept of credit is never defined in a principled way and so it is unclear what each algorithm is actually trying to estimate. This leads advertisers to use attribution results to solve a myriad of problems. For example, an advertiser may use last-click attribution from a set of data to evaluate the value of an entire ad campaign, decide how to allocate spend across multiple advertising campaigns or even figure out what to bid in an ad auction. By

not having a strict definition of what an attribution algorithm is trying to achieve, we are left with an algorithm that is used to solve multiple problems, while not being suitable to solve any one of them.

In addition to this ambiguity in the measurement objective, there are additional problems related to entrenched reporting practices. Advertisers want an automated and continuous method of measurement with unequivocal results that are easy to interpret. These expectations have lead to the establishment of requirements for attribution that imply underlying assumptions that do not hold in practice and, in fact, undermine the task of measuring ad effectiveness. These include the requirements:

- credit must add to the total number of reported conversions, both at the path level and in aggregate,
- credit is non-negative at the event level, and therefore also at the channel level,
- credit is additive and no credit is explicitly assigned to ad interactions,
- credit is not assigned to events in non-converting paths.

These restrictions reflect the desire from advertisers to have a simple measure of how their advertising works, but accommodating this desire leads to implicit assumptions that are limiting and unlikely to hold in practice. For example, by only considering non-negative credits, attribution models are implicitly assuming that advertising cannot have a negative effect. This is reasonable since the goal of advertising is to drive conversions. However, there can be instances in which an ad may cause a user not to buy a product, and if this is the case, this is a very strong and useful piece of information that is lost under existing attribution measurement. This non-negative constraint can also cause issues in the estimation process when advertising is completely, or close to being, ineffective. Due to the probabilistic nature of user behavior, having some negative credits is appropriate and expected, even if the overall ad campaign had a positive effect in expectation. Prohibiting negative credit from occurring at all places an unreasonable constraint on the estimation and can lead to bias.

Continuing, the fact that attribution methods attribute a credit to each individual event, rather than assigning credits to combinations or sequences of events, also indicates that an independence assumption is being made. That is, the sum of marginal effects are guaranteed to equal the joint effect of all events only if the underlying events are independent. This is quite unlikely in practice, especially if the notion of an advertising funnel holds true.

Lastly, the notion that it is necessary to account for every conversion both at the path and aggregate level can cause serious issues. Namely, it implies that all relevant events and pieces of information that went into the decision for the user to convert is completely observed and present in the user's path. That is, if unpaid (non-ad) events are present in the path then the assumption is that credit that can't be attributed to ads can be explained by these observed unpaid events. If only paid events are present then the assumption is that no external or user actions caused the user to convert other than the observed advertising interventions.

This requirement also ignores the user's propensity to convert. As an extreme example, consider the situation in which only paid ad events are in the path and these have no impact on user behavior. That is, a user converts for reasons not related to the ad and thus by definition the events preceding a conversion are independent of the conversion status. Reporting requirements dictate that the credit for the conversions will be divvied up among these events even when we know they had no effect. Data-driven methods should utilize all information in the data and, as is the case in Sapp and Vaver (2016), take the approach of considering non-converting paths in assigning credit. While this is a positive step for attribution, the goal of accounting for every conversion remains along with the problems associated with this expectation.

Beyond these issues with reporting, advertisers often choose an attribution model based

on preconceived notions about how attribution credit should be allocated, which can lead to poor decision-making. A better approach is to choose an attribution model based on its demonstrated ability to answer a specific decision-related question. The approach we take in this paper is to choose a single measurement goal, which is to estimate the number of additional conversions generated by a single ad channel. Here an ad channel is defined as a grouping of similar campaigns that the attribution algorithm treats as equivalent. This is inherently a causal question, so answering it requires defining core concepts such as credit in terms of a causal effect.

3 Attribution Through Experimentation

We take the viewpoint that attribution is actually trying to estimate the effectiveness of different types of advertising. Namely, to estimate the effect that an intervention (an ad) has on an outcome of interest (the conversion) and as such it is inherently a causal inference problem. The causal framework we outline is extremely important as it reflects the goal of attribution to comment on the effects of actions made by advertisers not merely to find which events are correlated with conversions.

Experimentation is the gold standard in causal inference, so attribution is best understood when viewed as a broken randomized experiment. That is, let us imagine what experiment we would do to estimate the effect of advertising and then let's examine how the observed data from attribution products fit into that experiment. This approach makes it clear when, and how, observational methods of measurement, such as attribution models, are appropriate. It requires the clear definition of an estimand and makes it possible to conduct simulation studies, as described in Section 7, that can be used to evaluate attribution models with experiments that are run on simulated path data. Most importantly, this causal motivation guides the development of new attribution methods that explicitly resolve discrepancies between the assumptions of the experiment and the attribution model.

Advertising is often organized such that multiple ad channels are concurrently active. This proposal is tackling one of the goals of attribution; determine how valuable each channel is in affecting a user's propensity to convert. It is hypothesized that for each ad channel there is a corresponding experiment of interest where in one arm of the experiment the ad channel is active and in the other arm it is inactive. The attributable value of the ad channel is defined as the difference between the number of conversions in the two experimental arms. Note that this marginal effect of an ad channel is still relevant when there are multiple ad channels. Each ad channel can be active or inactive and therefore has its own marginal effect even in the presence of other active channels. Additional effects can be defined with a more complex experiment in which multiple ad channels are concurrently active or inactive. This would allow for measures of interactions between ad channels.

One such experiment would be a full factorial experiment in which every combination of the ad channel is active or inactive. In Appendix A, we present a proposal that describes a set of contrasts that measure the relative importance of each channel across the full factorial experiment by combining main and interaction effects. This approach is very much in line with that proposed in Shapley (1953) and, although it partitions credit for the aggregate incremental conversions generated by the ad channels, it provides a different set of information to the advertiser that is arguably less actionable. This is because the incremental effect of a channel is directly tied to an advertiser action of turning a channel on or off. A Shapley value or the set of contrasts from a full factorial embedding multiple effects is, by definition, a measure under many different advertiser actions. Each combination of advertiser actions represents a particular arm of the full factorial experiment and, as only one arm of the experiment or set of advertiser actions can be taken at any one time, this makes the information much less actionable. On the other hand, this full factorial approach does accomplish the

goal of providing a systematic objective for allocating conversion credit across paid ad channels along with a no-advertising conversion baseline.

We tackle one use case for attribution; the effect of turning a whole channel on or off. This use-case is in line with estimating the return on ad spend (ROAS) at the channel level, which can advise the advertiser regarding the worth of an entire channel. Unlike existing attribution methods that assign credit to every event, the proposal addressed in this paper does not estimate the effect of each ad instance at the ad serving level. Identifying a credit for each ad event would be more useful for practices such as bidding on ad serving opportunities in an auction environment. However, these credits may not accurately estimate the ROAS of an entire channel. The fact that current attribution methods often try to answer these two questions simultaneously with one algorithm is a concern as equivalence exists only under some strict assumptions. Avoiding this pitfall is the first step in the systematic design of an attribution model. Although in this paper we only consider the ROAS case both questions can be formulated and answered within the causal framework proposed.

4 Causal Framework

As we indicated previously, attribution estimates the impact that an intervention has on an outcome and so it is natural to view it under the Rubin Causal Model (RCM) framework (Rubin, 1974). Under the RCM framework we can quantify the causal effect (the estimand), describe how to estimate this quantity (the estimator) and list assumptions such that the estimator has desirable properties such as unbiasedness.

An RCM aims to embed an observational study into a hypothetical randomized experiment that happens to be *broken* in that the exact assignment mechanism of the treatment, in this case ad exposure, is unknown to the researcher. This framework also allows us to imagine what value the outcome metric would have taken for each experiment arm. While this framework can be used to describe a multi-arm experiment with each arm having a single channel turned off, for exposition purposes, we focus on turning on and off a single channel in the presence of other channels that always remain on. This experiment aims to replicate the behavior and interests of advertisers who have multiple channels running on an ongoing basis and would like to know the marginal effect of a single channel in the presence of others. Knowing how a channel contributes to the number of conversions can inform the marginal return on ad spend (ROAS) for that channel.

For the attribution problem, the experimental units are individual users who have the potential to be served ads. In order to make the problem tractable, and for our inferences to be causal, it is necessary to make assumptions about the underlying data generation process. The first assumption we make is related to how the users and their corresponding digital histories are sampled:

Assumption 1 ***(Simple Random Sample from a Superpopulation)***
N users are independently and identically sampled from an infinite superpopulation. This is equivalent to setting an observation window and observing all the users with events in the window where the start and end dates of the window are independent of the distribution of the collected data.

The second assumption is about the ad serving mechanism and it is made by the majority of existing attribution models (often implicitly), including the rules-based ones:

Assumption 2 ***(Stable Unit Treatment Value Assumption - SUTVA)***
There are two components to this assumption. The first is that there is no interference between users, and the potential outcomes of one user are not affected by the treatment assignment of another. That is, whether or not one user sees an ad has no influence on whether a different user converts. Secondly, we assume that there are no hidden variations of treatment and that neither the label of treatment or the assignment has an effect on the potential outcomes. (Imbens and Rubin, 2015)

Under these first two assumptions we can equate the advertiser level experiment of turning off an entire channel with the individual effect of an ad at the user level. One consequence is that an advertising channel can only affect a user if that user actually sees an ad. This means that a user will have the same number of conversions if they do not see the ad when the channel is active as they would have if the channel was inactive. This identification allows us to estimate the effect of turning off an ad channel solely by observing those users who did and did not see ads from the channel when it was active.

While it is likely that the first two assumptions are required by most attribution models our third assumption could be relaxed and replaced by an alternate modeling assumption.

Assumption 3 ***(Conditionally the Passage of Time has No Effect)***
Given the sequence of events, whether they be ads or organic actions by the user, the time between events is independent of conversion status. That is, the actual calendar time of the events does not matter as all of the relevant information is contained in the sequence of events.

The consequence of Assumption 3 is that we only need to consider the sequence of events when doing attribution and not the timestamps of those said events. We assume this largely for expositional reasons and the theory outlined here can be expanded to incorporate temporal information.

Our fourth assumption is done so for mathematical reasons and is largely inconsequential as its veracity is confirmed for any problem in practice.

Assumption 4 ***(There are a maximum of J finite number of events in a user's path)***
A user can see an ad for the first time at any position in the sequence of events in their path, which is collected across a common observation window. We assume that each user path contains a finite number of events that is less than some integer, J.

From Assumption 4 we see that every user has the potential to see an ad for the first time at any position in their path or, alternatively, not at all. Thus, depending on when a user sees the ad, they may have a different number of conversions at the end of the experiment or observation window. These are known as potential outcomes under the RCM framework. Namely, for each user there is at a maximum, $J+1$ potential outcomes (number of conversions), depending if the user saw the ad and in what position they first saw it. Note that here we are implicitly assuming that each user has a fixed organic path that only changes through the intervention of an ad and does not change through repeated experiments. This allows us to treat the organic path as a covariate that is only partially observed depending on what position the ad was served. We do, however, allow for the effect of the advertising to be stochastic and hence the $J+1$ potential outcomes are random variables with each realization representing the number of conversions resulting from a realized path.

For every user (and ignoring any user index for now) we let $\{C^j \mid \text{ for } j = 0, 1, \ldots, J\}$ be the set of these potential outcomes (Splawa-Neyman et al., 1990; Rubin, 2005) where C^0 is defined as the number of organic conversions (when the user does not see the ad at all). We also let $Z \in \{0, 1, \ldots, J\}$ be the treatment index that identifies the position in a user's path in which they saw the ad for the first time. Finally, we let $\{U_j \mid \text{ for } j = 0, 1, \ldots, J\}$ be the subsequence of events that occurred upstream of, but not including, the first ad event at position j.

5 Estimand

Under Assumption 2, if a user does not see an ad when the channel is active then this user should have the same number of conversions as if the channel had been inactive. Hence, we can write

our estimand as

$$\begin{aligned}\Delta &= \mathrm{E}(\#\text{ Conversions when channel is active}) \\ &\quad - \mathrm{E}(\#\text{ Conversions when channel is inactive}) \\ &= \mathrm{E}(\sum_{j=1}^{J} \mathbf{1}_Z(j)(C^j - C^0)) \\ &= \sum_{j=1}^{J} \mathrm{E}(C^j - C^0 | Z = j)\mathrm{P}(Z = j) \\ &= \sum_{j=1}^{J} \Delta_j p_j \qquad (1)\end{aligned}$$

where $\Delta_j = \mathrm{E}(C^j - C^0|Z = j)$ and $p_j = \mathrm{P}(Z = j)$. Note that in the simple binary treatment case $\Delta_j = \mathrm{E}(C^j - C^0|Z = j)$ is a standard estimand known as the Average Treatment Effect for the Treated (ATT). However, in this case, there is an ATT for each event number, $\{\Delta_j \mid j = 1, \ldots, J\}$, in which a user could see the ad for the first time. Our estimand, Δ, then is just the average of these ATT's weighted by the probability of seeing the ad for the first time at each event number.

The issue with any experiment is that only one potential outcome can ever be observed for each user. That is, it's not possible to apply treatment (i.e., show an ad for the first time) at two different points in the path for any given user, let alone apply treatment for the first time at every point in the path for every user. This is known as the fundamental problem of causal inference (Holland, 1986) and it means that, for each user in our experiment, we will be missing all but one of $\{C^j \mid \text{ for } j = 0, 1, \ldots, J\}$. Note we defined Z as the variable denoting the point in the path that the ad was served in the realization of the experiment. The definition of Z along with the following identifiability assumption formally links our potential outcomes under seeing the ad at any point in the event sequence and the actual outcome that was observed in the experiment that took place,

Corollary 1 ***(Identifiability)***

$$C^{obs} = \sum_{j=0}^{J} C^j \mathbf{1}_Z(j).$$

A consequence of SUTVA (Assumption 2) is that the potential outcomes are statistically identifiable because any change in their distribution will naturally change the distribution of the observed data.

Our final assumption, Assumption 5, relates to the ad serving mechanism, otherwise known as the treatment assignment. Here we assume that by event number j, we have observed all of the relevant information related to whether a user may see an ad or not at that point so that the treatment status is independent of the value of the potential outcomes.

Assumption 5 ***(Strongly Ignorable Treatment Assignment)***
For $j = 1, \ldots, J$ assume

$$(C^0, C^j) \perp\!\!\!\perp \mathbf{1}_Z(j)|U_j$$

and

$$0 < P(Z = j|U_j = u) < 1$$

for all u and $j = 1, \ldots, J$.

A consequence of this assumption is that there are no unobserved confounders. That is, there are no missing variables that are related both to treatment and the potential number of conversions. This assumption is required in order to correctly identify and estimate the causal effect of advertising. If there was such a confounder it can easily be seen that any differences in conversion rates between users who saw ads and those who did not could, perhaps, be partially or fully explained by the missing confounder.

For our experiment this assumption equates to saying that the sequence of events upstream of the ad serving opportunity are all that is required to estimate the effect of the ad at that point in the event sequence. In general, this assumption does not hold due to the existence of ad targeting. It obviously ignores important covariate information, such as age and location, which are very often used in ad targeting. In order to align with common practices, and for our comparison to be on an even keel with rules-based attribution algorithms, we restrict ourselves to only having the sequence of events available.

However, making covariate information available is the biggest opportunity for improving attribution, and the algorithm we describe is easily extended to a use case in which additional covariate information is available.

6 Estimation

Our estimator is directly informed from the estimand in (1) as we can estimate Δ via a plug-in estimator by estimating each Δ_j and p_j separately for all j. This is desirable as we have reduced the rather complicated problem of attribution to estimating the average treatment effect on the treated (ATT) for a binary treatment, which is a standard problem.

Δ is the average effect for a randomly sampled user and hence we do not index by user in defining Δ. In our estimation, however, we utilize all experimental units (users) in our experiment and we use i to index these users.

To estimate p_j we simply use the sample mean in each group

$$\hat{p_j} = \frac{N_j}{N}$$

where $N_j = \sum_{i=1}^{N} \mathbf{1}_{Z_i}(j)$, and N is the number of users sampled as defined in Assumption 1.

In order to estimate Δ_j we need to further link our potential outcomes to the observed data. By the law of total expectation we see that

$$\begin{aligned}\Delta_j &= E(E(C^j - C^0|Z = j, U_j)|Z = j) \\ &= E(\Delta_{(j,U_j)}|Z = j)\end{aligned}$$

where $\Delta_{(j,u)} = E(C^j - C^0|Z = j, U_j = u)$. Then using (Corollary 1) and (Assumption 5) we find that

$$\begin{aligned}\Delta_{(j,u)} = &E(C^{obs}|Z = j, U_j = u) \\ &- E(C^{obs}|Z = 0, U_j = u). \qquad (2)\end{aligned}$$

(2) shows that $\Delta_{(j,u)}$ can be expressed as the difference in the mean number of conversions for users with upstream path, u, who saw the ad at event j and those users with the same upstream path, u, who did not see the ad at event j. This immediately leads to the following estimator,

$$\hat{\Delta}_{(j,u)} = \frac{C(j,u)}{N(j,u)} - \frac{C(0,u)}{N(0,u)}, \qquad (3)$$

where $C(j,u) = \sum_{i=1}^{N} \mathbf{1}_{Z_i}(j)\mathbf{1}_{U_{ij}}(u)C_i^{obs}$ represents the number of observed conversions and $N(j,u) = \sum_{i=1}^{N} \mathbf{1}_{Z_i}(j)\mathbf{1}_{U_{ij}}$ the number of users who saw an ad in position j with upstream path u.

We can interpret $\hat{\Delta}_{(j,u)}$ as the difference between sample conversion rates for users with upstream path u who saw the ad at time j and those who did not. In short it is an estimate of the incremental effect of advertising at time j on users who have already taken path u. For the j^{th} event, there are many different possible upstream paths. We can average over these upstream paths to estimate the effect of advertising for this event number,

$$\hat{\Delta}_{(j)} = \sum_{u} \hat{\Delta}_{(j,u)} \frac{N(j,u)}{N_j}. \qquad (4)$$

We now have an estimate for the effect of advertising for each value of j and we can average over the distribution of when users saw the ad in order to estimate the overall average effect of advertising in the treated population,

$$\hat{\Delta} = \sum_{j=1}^{J} \hat{\Delta}_{(j)} \hat{p}_j. \qquad (5)$$

Note that this model is equivalent to the one described in Sapp and Vaver (2016).

To demonstrate how this estimator works we revisit the path from Figure 1. In this path, we note that the user saw the *Search Ad Impression*[1] in the third position in the path, $Z = 3$, and had upstream path $U = (\textit{Direct Navigation}, \textit{Display Impression})$. In Figure 2, we have

[1] A search ad impression represents the intervention taken by the advertiser. We are interested in the causal effect of this impression but attribution models often only have access to search clicks. As such, under some assumptions about the click through rate, search clicks are used in practice as proxies for impressions even though they are outcomes and are caused by the intervention.

aggregated all those paths that have this same upstream path, and then separated them into groups that saw an ad in position 3 and those that did not see an ad at all. From (3) we construct the estimator,

$$\hat{\Delta}_{3,U} = \frac{10}{100} - \frac{20}{500} = 0.1 - 0.04 = 0.06.$$

To estimate the effect of the entire ad channel for search, Δ, this process is repeated for all uniquely observed upstream paths to estimate Δ_3, and then this process is repeated to find each Δ_j.

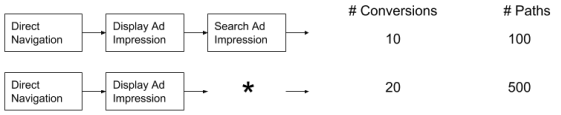

Figure 2: An Example of Estimation

6.1 Missing Data Bias Adjustment

The estimator in (5) would be sufficient if it weren't for a data collection issue that plagues attribution products. By definition the data collection system only collects user paths that have at least one observable event within the observation window. Some events are not detectable by the data collection process and so users only enter the database based on whether they have an observable event in the observation window. This issue is described in Sapp and Vaver (2016) in the section *Systematically Censored Users* where the authors note the bias of upstream data driven attribution in the presence of this censoring.

This censoring would not be an issue if the censoring affected both treated (those who see an ad) and control (those who do not see an ad) users the same. However, given that seeing an ad is captured by the data collection mechanism we see that all treated users are observed and the control users, who do not have an observable event, will be missing. Hence, using the estimator described in (4) and (5) naively without adjusting for this censoring will result in a biased estimate of the causal effect.

This problem only arises in the estimation of Δ_1. This is because users who do not immediately see an ad, i.e., users with $j \geq 2$, we know that the first event in the path must have been an observable non-ad event. In the estimation of Δ_j the data collection process affects both treated and control equally as all units entered the sample due to the first observable event. On the other hand, estimating Δ_1 is problematic because there is no prior observable event to match on when constructing a counterfactual control group. All of the relevant treated users are observed by virtue of seeing the ad, however, not all users who could have seen the ad, but didn't, are observed as some may not have any observable event in the observation window.

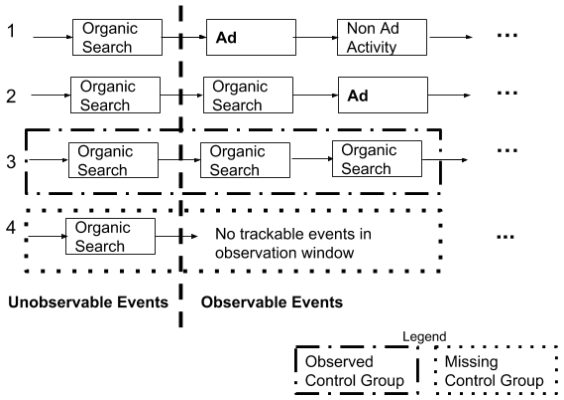

Figure 3: Missing Data Example

This issue is demonstrated in Figure 3 where both *User 1* and *User 4* have identical upstream paths leading up to the start of the observation window. *User 4* should be used in the estimation of the counterfactual control conversion rate for *User 1* but as our observation window did not encompass any observable events for *User 4* they were censored and did not enter the dataset. The same issue does not exist in estimating the counterfactual conversion rate for *User 2*, however, as there is an observable first event and hence the counterfactual control path (*User 3*) will also have an observable first event and enter the dataset.

Since we assumed that the observation window is selected at random, and given our assumption that only the sequence of events are informative,

rather than the passage of time (Assumption 3), the distribution of upstream paths leading to an ad should be the same just prior to the start of the observation window as it is in the observation window. Hence, the group of users who immediately see an ad can be thought of as a mixture of the treated users, each with varying upstream paths or, equivalently, as a random sample of treated users who happened to have their upstream path censored. This is illustrated in

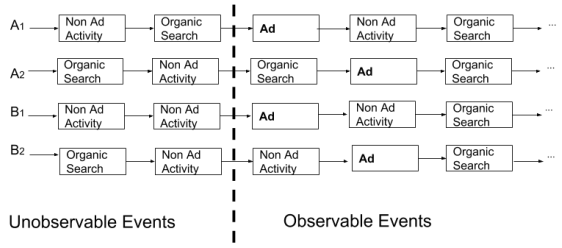

Figure 4: Mixture Example

Figure 4 where *User* A_1 and *User* B_1 are indistinguishable in the observation window. If the window had begun earlier, we would be able to distinguish them as is possible with *User* A_2 and *User* B_2. In this example we see that the distribution of the missing upstream paths is the same as the observed upstream paths in the observation window.

This property is implied by the assumption that the passage of time is uninformative given the event sequence. Under this assumption, and the fact that the missing upstream paths would have the same distribution as the observed ones in the observation window, we can construct the estimator for Δ_1,

$$\hat{\Delta}_1 = \frac{\sum_{j=2}^{J} \hat{\Delta}_{(j)} \hat{p}_j}{\sum_{j=2}^{J} \hat{p}_j}, \tag{6}$$

which can be used to compute the first term in (5). Note that (6) is simply the average effect of seeing the ad at each position weighted by the estimated probability of seeing the ad in that position. Or, equivalently, it is just the estimated average treatment effect for those who saw an ad after the first event. This adjustment is a fix to censoring issue for the *UDDA* model described in Sapp and Vaver (2016).

Note that we can test Assumption 3 which underpins the theory behind our calculation in (6). This is because, although we cannot construct a counterfactual for the group of users who immediately see an ad, we can compare the observed conversion rate under seeing the ad with the rate implied by the mixture. That is, we have two estimates of the conversion rate under seeing the ad, r^1, where $r^j = E(C^j|Z = j)$. The first estimate uses the sample rate for all users who happen to see the ad immediately, $\hat{r}^1$, where

$$\hat{r}^j = \frac{\sum_{i=1}^{N} C_i^{obs} \mathbf{1}_{Z_i}(j)}{N_j}.$$

and the second uses the rate implied by the mixture, $\sum_{j=2}^{J} \hat{r}^j \hat{p}_j$. We can test to see if these two estimates are truly estimating the same true population quantity by undergoing a two-sample t-test utilizing the point and standard error estimates of each method.

7 Evaluation

One way to evaluate the method described in Section 6 is to see how closely the attribution algorithm can estimate the true effect of advertising. This is accomplished by comparing estimates from the attribution algorithm with results from corresponding experiments undertaken on real users. However, this is typically a costly procedure and advertisers are often reluctant to turn off entire ad channels due to the missed opportunity cost of not serving ads to potential converting users. Hence, we strive to provide some level of validation through a simulation study in which we simulate the generation of user paths and the ways in which these paths are affected by ads.

To accomplish this we utilize the Digital Advertising System Simulation (DASS) (Sapp et al., 2016). DASS models the way in which users traverse the internet via a non-stationary Markov process. It generates user level path data and has the flexibility to do so under many different assumptions about how advertising affects user behavior. We use DASS to generate sets of path data for which we can find the true effect

of an ad channel by allowing simulated users to experience both arms of an experiment (ad channel is active/inactive). The difference between the number of conversions in these two arms is by definition, Δ, as defined in (1). This experimental ground truth is used to determine how well an algorithm fairs in estimating Δ when it only observes path data from the experimental arm in which the ad channel is active, as is the case in practice.

We compare the algorithm described above, *UDDA Bias Adjusted*, to the existing algorithm, *UDDA*, as described in Sapp and Vaver (2016) where *UDDA* stands for Upstream Data-Driven Attribution. Note that *UDDA Bias Adjusted* is functionally identical to *UDDA* without the bias adjustment step noted in Section 6.1.

It is important to evaluate an attribution algorithm under many different conditions (e.g., different types of ad impact on user behavior, different magnitudes of ad impact, and different mixes of ad channels) to help ensure that it has the flexibility to handle situations that are encountered in practice. See Singh and Vaver (2017) for a more in depth discussion of model evaluation and the development of evaluation scenarios. Here we focus on the evaluation scenarios previously considered by Sapp and Vaver (2016).

In Figure 5 we examine how *UDDA Bias Adjusted* compares with the *UDDA* algorithm introduced in Sapp and Vaver (2016). This scenario examines the ability of an attribution algorithm to identify changes in the effectiveness of *Display* advertising. In this scenario, each time a user sees a display ad impression there is a chance that the user's subsequent browsing behavior will be altered. In moving from left to right on the x-axis, the effectiveness of the ad is increased in the sense that users are more likely to favor activities that will lead to a conversion (e.g., do an advertiser-related search or visit the advertiser's website). The *Truth* line indicates the true difference in the number of conversions between the two experimental arms in which the *Display* ad channel is active versus inactive. The error bars on the ground truth represents a 95% interval for the ground truth constructed through repeated experiments.

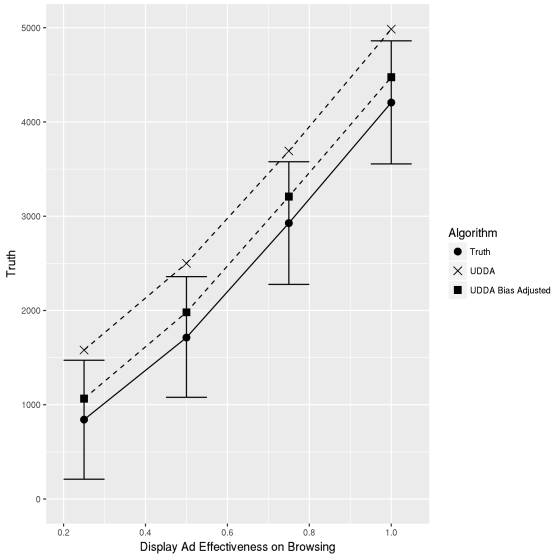

Figure 5: This plot is analogous to the first plot from Figure 6 in Sapp and Vaver (2016) in which display impressions affect downstream browsing behavior.

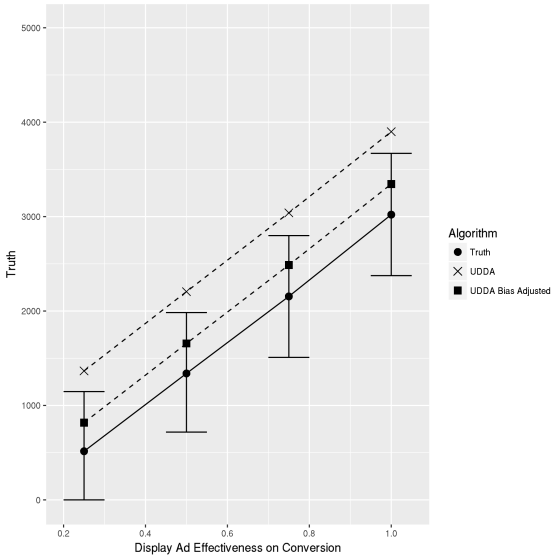

Figure 6: This plot is analogous to the second plot in Figure 6 from Sapp and Vaver (2016) in which display impressions affect the probability of conversion.

Figure 6 has a similar setup to Figure 5 with the exception that in this scenario the display advertising only affects a user's propensity to convert and does not otherwise alter the user's downstream browsing behavior. Again, in this

scenario we see that *UDDA Bias Adjusted* closely matches the *Truth*.

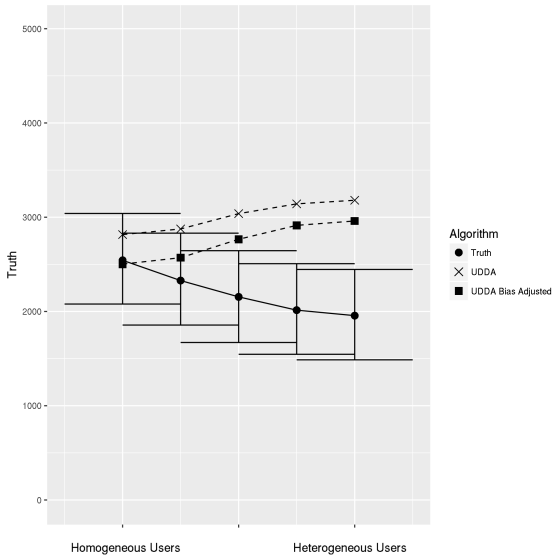

Figure 7: Here we visit a scenario in which the ad serving mechanism targets users differently depending on some covariate information not available to the attribution algorithm. In moving along the x-axis from left to right the effect of ad targeting diminishes to zero as the users become more homogeneous and hence are more similarly targeted.

Figure 7 corresponds to a scenario in which both *UDDA* and *UDDA Bias Adjusted* should have trouble. Poor performance is expected because there is extraneous covariate information not available to the attribution algorithms that is needed to inform both a user's propensity to see an ad and their propensity to convert. This is a violation of Assumption 5. The result is that counterfactual comparisons may be comparing groups with different distributions of important covariates that explain some of the difference in conversion rates.

As expected, *UDDA Bias Adjusted* fairs worse when ad targeting is more prevalent and is unbiased when there is no form of ad targeting. This is a problem that can't be fixed by changing the attribution algorithm. It can only be fixed by providing the attribution algorithm with covariate data related to ad targeting.

8 Conclusions

In digital advertising, attribution is widely used to inform a variety of marketing decisions. In this paper we highlight the need to identify a singular objective (Section 2) for attribution measurement and propose that advertisers are really interested in the marginal effect of an ad channel.

In viewing attribution through a scientific lens, we propose a hypothetical experiment in Section 3 that attribution products might aim to approximate. This experimental viewpoint allows for the creation of an easily interpretable estimand of interest (Section 5) which is based on a replicable action by the advertiser (turning off an ad channel).

In Section 4 we describe the attribution problem in a causal framework using the Rubin Causal Model (RCM). The benefit is that this approach requires us to construct the assumptions that are needed to be able to identify and estimate the estimand (causal effect) of interest. These assumptions are made so that we can more easily identify situations in which an attribution algorithm will do well or fall short in terms of estimating ground truth. They also make it easier to identify avenues for algorithm or data source improvement. The specific description provided was largely made for the purpose of demonstrating how a simple attribution algorithm can fit into a causal framework. In practice, and depending on the particular attribution environment, some of these assumptions may not hold and can be removed, relaxed or extended.

In Section 6.1, we propose a solution for the problem of working with path data that systematically censors users. This issue has plagued previous attribution algorithms and left them unusable in practice (Sapp and Vaver, 2016).

In Section 7 we conclude with a simulation study highlighting the efficacy of the *UDDA Bias Adjusted* algorithm. We use the DASS simulator to compare the estimates of *UDDA* and *UDDA Bias Adjusted* to the *Truth* and note the vast improvement over *UDDA* due to the missing data adjustment. In Figures 5 and 6, we see that *UDDA Bias Adjusted* improves *UDDA* and now all estimates fall within the 95% interval. Addi-

tionally, in Figure 7 we see improvements when ad targeting is present and hence there is a large degree of bias and estimation is difficult. We conclude that *UDDA Bias Adjusted* is a viable solution to the *Systematically Censored Users* problem described in Sapp and Vaver (2016).

In short, we have highlighted issues with current attribution practices and algorithms and described the attribution problem under a causal framework. This framework provides a useful guide for understanding and addressing deficiencies in attribution modeling. It should be the basis for all future attribution model development efforts.

9 Acknowledgments

We would like to thank many of our colleagues at Google for their comments and suggestions. In particular we acknowledge the attribution team in Advanced Measurement Technologies for their many discussions and Tony Fagan and Amy Richardson for providing comments and feedback.

A Proposal for Embedding Synergistic Effects Into One Measure Per Channel

A.1 Model for two channels

Consider the case of two channels, A and B, where we can run a full-factorial experiment as shown in Figure A1. One model formulation is as follows: [2]

$$Y_{ij} = \beta_0 + \beta_1 \mathbf{1}_i(1) + \beta_2 \mathbf{1}_j(1) + \beta_{12} \mathbf{1}_i(1)\mathbf{1}_j(1) + \epsilon_{ij},$$

where the i subscript is for channel A indicating off or on by 0 or 1, respectively, and the j subscript is for channel B similarly indicating off and on by 0 and 1. This can be written in matrix notation as $Y = X\beta + \epsilon$ where Y is a column vector given in the first column of Table A1, X is a matrix given in the last four columns of Table A1, X is a matrix given in the last four columns of Table A1, $\beta = (\beta_0, \beta_1, \beta_2, \beta_{12})^T$ and ϵ is a vector of error terms. [3]

	A	B	X_0	X_1	X_2	X_{12}
Y_{00}	0	0	1	0	0	0
Y_{10}	1	0	1	1	0	0
Y_{01}	0	1	1	0	1	0
Y_{11}	1	1	1	1	1	1

Table A1: Full factorial experiment for two channels along with corresponding design matrix X (last four columns)

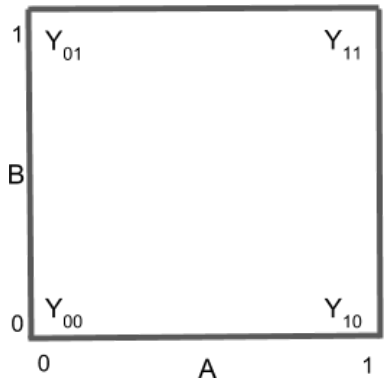

Figure A1: Full factorial experiment for two channels A and B with resulting outcomes Y_{ij}

We can solve for β using least squares $\hat{\beta} = (X^T X)^{-1} X^T Y$ where the matrix $Q = (X^T X)^{-1} X^T$ is the linear combinations of the data used to estimate the parameters of the model. For the two channel case this is given in Table A2. So, for example, $\hat{\beta}_0 = Y_{00}$ and $\hat{\beta}_{12} = (Y_{00} - Y_{10} - Y_{01} + Y_{11})$.

Now consider our estimate for the total number of incremental conversions:

$$\Delta C = Y_{11} - Y_{00} = \beta_1 + \beta_2 + \beta_{12}$$

[2] This deviates from the classic analysis of experiments model where the intercept represents all channels are half on and half off and where the X-variables are coded as -1 and +1. However, the proposed formulation provides easier understanding of the de-duping of credit and how to adjust to fit into the ROAS reporting requirements.

[3] The error terms aren't particularly interesting here as the rank of the model equals the number of observations so estimation of the errors can't be derived from the model - although they could be from external sources (i.e., estimated from a simulator directly). We will assume that the errors are negligible.

	Y_{00}	Y_{10}	Y_{01}	Y_{11}
0	1	0	0	0
1	-1	1	0	0
2	-1	0	1	0
12	1	-1	-1	1

Table A2: Q matrix for two channels.

and the estimates of the marginal impact of each channel (e.g., impact of the target channel when the other channel is on):

$$\Delta C(A; B \text{ on}) = Y_{11} - Y_{01} = \beta_1 + \beta_{12}$$
$$\Delta C(B; A \text{ on}) = Y_{11} - Y_{10} = \beta_2 + \beta_{12}.$$

Clearly, the inconsistency of the marginal impact estimates and the desire that these estimates sum to the total ad-driven conversions is the double counting of the synergistic or interaction effects between the two channels A and B (i.e., β_{12}). A reasonable solution is to give equal credit to each channel for the interaction effects. Hence,

$$\Delta A = \beta_1 + \beta_{12}/2$$
$$\Delta B = \beta_2 + \beta_{12}/2$$

which gives the desired property that $\Delta A + \Delta B = \Delta C$. These adjusted credits can be easily calculated as linear combinations of the Y's:

$$\begin{aligned}\Delta A &\approx (Q_{2\cdot} + Q_{4\cdot}/2)Y \\ &= (-0.5, +0.5, -0.5, +0.5)Y \\ &= 0.5(Y_{10} - Y_{00}) + 0.5(Y_{11} - Y_{01}) \\ \Delta B &\approx (Q_{2\cdot} + Q_{4\cdot}/2)Y \\ &= (-0.5, -0.5, +0.5, +0.5)Y \\ &= 0.5(Y_{01} - Y_{00}) + 0.5(Y_{11} - Y_{10}).\end{aligned}$$

So now the estimates are the average of the effect of turning a channel on when the other channel is on and is off. Referring to Figure A1, this is for channel A the average effect of moving from the left side of the square to the right side and for channel B moving from bottom to top.

A.2 Model for Three Channels

The model for two channels extends easily to three channels. Here we denote the three channels as A, B, and D[4] and now use subscript k for channel D - again 0 indicating the channel is off and 1 indicating the channel is on. The model is

	β_0	β_1	β_2	β_3	β_{12}	β_{13}	β_{23}	β_{123}
Y_{000}	1	0	0	0	0	0	0	0
Y_{100}	1	1	0	0	0	0	0	0
Y_{010}	1	0	1	0	0	0	0	0
Y_{110}	1	1	1	0	1	0	0	0
Y_{001}	1	0	0	1	0	0	0	0
Y_{101}	1	1	0	1	0	1	0	0
Y_{011}	1	0	1	1	0	0	1	0
Y_{111}	1	1	1	1	1	1	1	1

Table A3: Full factorial design matrix, X, for three channels.

$$\begin{aligned}Y_{ijk} = \beta_0 &+ \beta_1 \mathbf{1}_i(1) + \beta_2 \mathbf{1}_j(1) + \beta_3 \mathbf{1}_k(1) \\ &+ \beta_{12} \mathbf{1}_i(1)\mathbf{1}_j(1) + \beta_{13} \mathbf{1}_i(1)\mathbf{1}_k(1) \\ &+ \beta_{23} \mathbf{1}_j(1)\mathbf{1}_k(1) + \beta_{123} \mathbf{1}_i(1)\mathbf{1}_j(1)\mathbf{1}_k(1) \\ &+ \epsilon_{ijk} \qquad (7)\end{aligned}$$

Again, this can be written in matrix notation as $Y = X\beta$ where the X matrix is given in Table A3. The visualization of the design is given in Figure A2 and similar to the two channel case $\hat{\beta} = (X^T X)^{-1} X^T Y = QY$ and the Q matrix is given in Table A4.

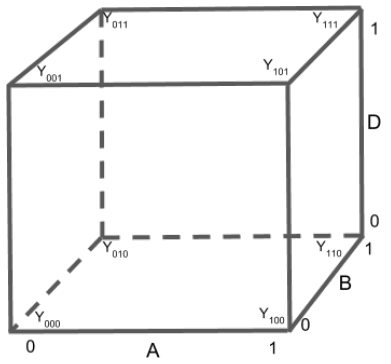

Figure A2: Full factorial experiment for three channels A, B and D with resulting outcomes Y_{ijk}

The estimate for the total number of conver-

[4]We use D to not confuse C with conversions.

	Y_{000}	Y_{100}	Y_{010}	Y_{110}	Y_{001}	Y_{101}	Y_{011}	Y_{111}
β_0	1	0	0	0	0	0	0	0
β_1	-1	1	0	0	0	0	0	0
β_2	-1	0	1	0	0	0	0	0
β_3	-1	0	0	0	1	0	0	0
β_{12}	1	-1	-1	1	0	0	0	0
β_{13}	1	-1	0	0	-1	1	0	0
β_{23}	1	0	-1	0	-1	0	1	0
β_{123}	-1	1	1	-1	1	-1	-1	1

Table A4: Q matrix for three channels.

sions is

$$\begin{aligned}\Delta C &= Y_{111} - Y_{000} \\ &= \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123}\end{aligned}$$

and for the marginal impact of each channel (i.e., impact of the target channel when the other channels are on) it is

$$\begin{aligned}\Delta C(A; B \text{ and } D \text{ on}) &= Y_{111} - Y_{011} \\ &= \beta_1 + \beta_{12} + \beta_{13} + \beta_{123} \\ \Delta C(B; A \text{ and } D \text{ on}) &= Y_{111} - Y_{101} \\ &= \beta_2 + \beta_{12} + \beta_{23} + \beta_{123} \\ \Delta C(D; A \text{ and } B \text{ on}) &= Y_{111} - Y_{110} \\ &= \beta_3 + \beta_{13} + \beta_{23} + \beta_{123}.\end{aligned}$$

We have the two-way interaction terms being double counted while the three-way interaction term is triple counted. Applying split credit across interacting channels gives our adjusted credits

$$\begin{aligned}\Delta A &= \beta_1 + \beta_{12}/2 + \beta_{13}/2 + \beta_{123}/3 \\ &\approx (Q_{2\cdot} + Q_{5\cdot}/2 + Q_{6\cdot}/2 + Q_{8\cdot}/3)Y \\ \Delta B &= \beta_2 + \beta_{12}/2 + \beta_{23}/2 + \beta_{123}/3 \\ &\approx (Q_{3\cdot} + Q_{5\cdot}/2 + Q_{7\cdot}/2 + Q_{8\cdot}/3)Y \\ \Delta D &= \beta_3 + \beta_{13}/2 + \beta_{23}/2 + \beta_{123}/3 \\ &\approx (Q_{4\cdot} + Q_{6\cdot}/2 + Q_{7\cdot}/2 + Q_{8\cdot}/3)Y.\end{aligned}$$

We get the desired property that $\Delta A + \Delta B + \Delta D = \Delta C$. The detailed estimating coefficients are given in Table A5. All of these estimates are weighted contrasts from when a channel is on vs. off. For the three channel case, turning on a channel when either all other channels are off or all channels are on are weighed twice as much as when only one other channel is on.

	Y_{000}	Y_{100}	Y_{010}	Y_{110}	Y_{001}	Y_{101}	Y_{011}	Y_{111}
Δ A	-1/3	1/3	-1/6	1/6	-1/6	1/6	-1/3	1/3
Δ B	-1/3	-1/6	1/3	1/6	-1/6	-1/3	1/6	1/3
Δ D	-1/3	-1/6	-1/6	-1/3	1/3	1/6	1/6	1/3

Table A5: Estimating coefficients for estimating adjusted (additive) credit for three channels.

A.3 Model for General Number of Channels

The model is extended to the general case with p channels. Now the indices denote channels rather than levels of channels as in the examples above. The full factorial design can be represented by the vertices of a p-dimensional hypercube - each vertex representing a combination of channels on and off. Hence, we have $n = 2^p$ simulation results. Let the n-length column vector x_i represent the on/off settings (0/1) for channel i. Further, define the n-length column vectors $x_{ij} = x_i x_j, x_{ijk} = x_i x_j x_k, \ldots, x_{12\ldots p} = x_1 x_2 \ldots x_p$. The model,

$$\begin{aligned}Y_{ij\ldots p} = \beta_0 &+ \sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} x_{ij} \\ &+ \sum_i \sum_{j<i} \sum_{k<j} \beta_{ijk} x_{ijk} + \ldots \\ &+ \beta_{12\ldots p} x_{12\ldots p} + \epsilon_{12\ldots p},\end{aligned}$$

can again be represented as $Y = X\beta$ where we can compute $Q = (X^T X)^{-1} X^T$ to get the linear combinations of the Y's that estimate the β's. The estimate of the total number of incremental conversions is

$$\begin{aligned}\Delta C = &\sum_i \beta_i + \sum_{i<j} \beta_{ij} \\ &+ \sum_i \sum_{j<i} \sum_{k<j} \beta_{ijk} + \ldots + \beta_{12\ldots p}.\end{aligned}$$

While the marginal impact of channel i (turning i^{th} channel on when all others are already on) is

$$\begin{aligned}\Delta(i;\ \text{rest on}) = \beta_i &+ \sum_{j\neq i} \beta_{ij} + \sum_{k<j, k\neq i} \beta_{ijk} \\ &+ \ldots + \beta_{12\ldots p}.\end{aligned}$$

All two-way interactions are double counted, three-way triple counted, ..., and the p-way interaction is p-times counted. Hence the adjusted credit for channel i is

$$\Delta(i) = \beta_i + \sum_{j \neq i} \beta_{ij}/2 + \sum_{k<j; k \neq i} \beta_{ijk}/3 + \ldots + \beta_{12\ldots p}/p.$$

These can be easily computed by taking the appropriate rows of the Q matrix, dividing each row by the appropriate de-duping and then adding to get the contrast to calculate $\Delta(i)$.

References

Analytics, G. (2017). Google analytics help center: About the default attribution models. Technical report, Google Inc. `https://support.google.com/analytics/answer/1665189`.

Analytics, G. (2018). Google attribution capabilities. Technical report, Google Inc. `https://www.google.com/analytics/attribution/capabilities/`.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Sapp, S. and Vaver, J. (2016). Toward improving digital attribution model accuracy. Technical report, Google Inc. `https://research.google.com/pubs/pub45766.html`.

Sapp, S., Vaver, J., Shi, M., and Bathia, N. (2016). Dass: Digital advertising system simulation. Technical report, Google Inc. `https://research.google.com/pubs/pub45331.html`.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Singh, K. and Vaver, J. (2017). Attribution model evaluation. Technical report, Google Inc.

Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.