CAUSALLY MOTIVATED ATTRIBUTION FOR ONLINE ADVERTISING - Working Paper

Brian Dalessandro
M6D Research, NYC,USA
briand@m6d.com

Ori Stitelman
M6D Research, NYC,USA
ori@m6d.com

Claudia Perlich
M6D Research, NYC,USA
claudia@m6d.com

Foster Provost
NYU & M6D Research, NYC,USA
fprovost@stern.nyu.edu

ABSTRACT

In many online advertising campaigns, multiple vendors, publishers or search engines (herein called channels) are contracted to serve advertisements to internet users on behalf of a client seeking specific types of conversion. In such campaigns, individual users are often served advertisements by more than one channel. The process of assigning conversion credit to the various channels is called “attribution,” and is a subject of intense interest in the industry. This paper presents a causally motivated methodology for conversion attribution in online advertising campaigns. We first propose a need for the standardization of attribution measurement and offer four principles upon which standardization may be based. Stemming from these standards, we offer an attribution solution that generalizes prior attribution work in cooperative game theory and recasts the prior work through the lens of a causal framework. We argue that in cases where causal assumptions are violated, our solution can be interpreted as a variable (or channel) importance measure. Finally, we present a practical solution towards managing the potential complexity of the generalized attribution methodology, and show examples of attribution measurement on several online advertising campaign data sets.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - data mining; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Performance, Theory

Keywords

Attribution, Digital Advertising, Causal Estimation, Variable Importance

1. INTRODUCTION

Total US internet advertising revenue is projected to approach $40bn by the end of 2012 [7]. Estimates suggest that the internet accounts for only 13-19% of total media ad spend, despite taking up 25% of total time spent on media. This discrepancy suggests a significant opportunity for growth in online ad spending. Online ad spend is largely dominated by two categories: Paid Search, which has the dominant share of money spent, and Display, which includes banner ads and emerging video formats. The prevailing view is that a barrier to the growth of Display is the lack of appropriate measurement techniques that capture the full value of the category [6]. The measurement standard most debated in the industry is conversion attribution, which is generally defined as the assignment of conversion credit when multiple advertising channels reach a given online user, where an “advertising channel” can be defined as any vendor or publisher showing advertisements on behalf of an advertiser.[1] The current default conversion attribution methodology used in the industry is generally referred to as “last-touch” attribution, which is a rule-based attribution methodology that assigns full conversion credit to the channel that last presented an advertisement to a converting user. An attribution solution that can move beyond the last-touch model is regarded as a priority within the industry.

An alternative to the last-touch attribution has been proposed and is used by several firms within the industry. This alternative is generally referred to as the “multi-touch” attribution model, and attempts to assign credit to multiple channels when more than one have been observed to show an advertisement to a converting user.[2] Several firms offer heuristic based attribution methods [11, 12, 13], which use different sets of general rules to allocate conversion credit. Shao et al. [20] propose a multi-touch model that is statistically derived from the data generated by a specific adver-

[1] Examples of advertising channels used in this context include, but are not limited to, search engines, ad networks, behavioral targeting firms and large online publishers

[2] To be clear, this applies to scenarios where an advertiser elects to serve the same ads through multiple channels on a given campaign

tising campaign, and is, to the authors' knowledge, the only published data-driven system in practice.

Despite the progress that has been made in developing multi-touch attribution systems for online advertising, there is currently no existing generally accepted framework or set of standards upon which attribution measurement is based. Lack of transparency and standardization are considered to be two general barriers to the growth in online ad spend [8]. Of particular concern is the lack of transparency and standardization in online attribution measurement. Our goal with this paper is to make progress in this direction. Our first contribution is a set of general properties we propose in an effort to standardize online conversion attribution measurement. Guided by these properties, as a second contribution, we position multi-touch attribution as a causal estimation problem and present a general model for multi-touch attribution measurement that generalizes prior attribution work in Cooperative Game Theory. Third, we establish that prior work in the field of multi-touch attribution for online advertising falls within our principled framework. Finally, we present a practical specification of this general formula that can be easily implemented in standard production settings.

This paper is organized as follows: Section 2 proposes four properties of a good attribution system and presents a brief survey of the variable importance and causal methodologies that inform our attribution system. Section 3 starts with a formalization of our generalized causal attribution model and establishes the connection between this model and prior work in Cooperative Game Theory. We also present in section 3 a practical specification of the general model and establish that this is a generalization of prior work (published in KDD last year). Section 4 covers estimation and gives an empirical analysis on both simulated and real campaign data.

2. AXIOMATIC PROPERTIES, VARIABLE IMPORTANCE AND CAUSALITY

In this section we propose four general properties of good attribution systems. These properties motivate a causal modeling approach towards attribution measurement. We thus follow the presentation of the four properties with an overview of a commonly used framework for causal modeling. Finally, we survey common approaches to variable importance measurement and discuss the connections between causal inference and variable importance measurement, particular as they apply to attribution measurement.

2.1 Properties of a Good Attribution System

A good attribution system, more than anything, should align the incentives of the advertiser with the incentives of the channel hired to run ads on behalf of the advertiser. A common goal of an advertiser running an online ad campaign is to drive as many conversion events at as low a cost as possible (often measured by Cost Per Action and/or Return on Investment). The goal for channels, however, is to receive as much payment as possible for the conversion events the advertiser observes while minimizing its own costs. With multiple channels on a campaign (as is often the case), the optimal strategy for a given channel might be guided more by how it earns credit for a conversion and less by how well it can influence a conversion. An example of this can be seen in the "last-touch"' attribution model defined above and currently in widespread use. Under this attribution scheme, a channel might choose to run a large volume of cheaper "below-the-fold" advertisements (meaning the ad is placed below the main viewing window and is rarely viewed by the web user). Because a below-the-fold ad is likely to be the last to render on a page, and because of the quantity of ads served, the channel executing this strategy is likely to be the "last-touch" and thus earn much of the credit for conversions actually driven by the other channels.

We propose the following four properties in an effort to design an attribution standard that better aligns the incentives of advertisers and channels.

1. Fairness - A good attribution system should reward an individual channel in accordance with its ability to affect the likelihood of conversion (with conversion defined however is relevant to the parties involved).

2. Data Driven - A good attribution system should be derived specifically for the advertising campaign in question, using both ad treatment and conversion data captured during the campaign.

3. Interpretability - A good attribution system should be generally accepted by all parties with material interest in the system, on the basis of its statistical merit, as well as on the basis of intuitive understanding of the components of the system.

4. Economic efficiency - A good attribution system should account for the total value created by an ad campaign without requiring arbitrary normalizations to do so.

2.2 Attribution and Causality

We satisfy "fairness"' in our attribution system by using the *counterfactual* framework presented by Rubin in [19]. A counterfactual analysis in the context of online advertising asks the question: "what is the effect of an ad on user conversion?" Answering this involves measuring the difference in conversion outcome on a user with and without an ad being shown. As a single user can not be both "exposed"' and "unexposed" to an ad, the effect is usually measured as the difference of means between a group of "exposed"' users and a group of "unexposed"' users. Assuming that users are randomly assigned to either group, this measure is an unbiased estimate of the true causal parameter.

We set up our counterfactual framework for multi-touch attribution measurement by defining parameters that represent various levels of ad exposure (these will be defined explicitly in section 3). Estimation of these parameters can be achieved via direct controlled experimentation (such as with A/B testing similar to that done by Lewis et al. in [15]), or with causal estimation on observational data, as is done by Stitelman et al. [22] and Chan et al. [4]. This paper focuses specifically on counterfactual analysis using observational data. Randomized experiments are an appropriate and popular method for causal estimation in many applications, but in the multi-channel setting, it is often technically and practically infeasible to coordinate a large scale randomized experiment across a multitude of competing channels. Observational methods for multi-touch attribution measurement therefore are needed.

In order for a counterfactual analysis to produce unbiased estimates of causal effects (both in observational and experimental settings), several strict assumptions must be made about the data. These are: (1) that the treatment precedes the outcome (SUTVA, or time-ordering assumption), (2) that any attribute that may affect both ad treatment and conversion outcome is observed and accounted for (no unmeasured confounding) and (3) that every user has some non-zero probability of receiving an ad treatment (positivity assumption) [24]. It is highly likely that in observational data for multi-channel advertising campaigns the second and third assumptions will be violated. This reality does not prevent the use of counterfactual analysis, but it does require a change in the interpretation of results. In prior work related to counterfactual analysis with similar violations of the strict causal assumptions, the problem has been recast as a variable importance measure [25]. We continue with this precedent and cast attribution as a variable (or channel) importance problem, that under certain conditions, can be interpreted as strict causal analysis.

2.3 Attribution and Variable Importance

In cases where multi-touch attribution measurement can not be cast and interpreted as a causal effect analysis, the literature on variable importance supports recasting the problem as a channel importance problem. The counterfactual concept is found throughout various diverse approaches to variable importance, which we review in this section.

Variable importance is often defined in the literature as the effect on a measurable quantity of interest upon changing a variable of interest, holding all other factors constant. This definition is consistent with the counterfactual framework presented above (i.e., a variable can be a channel ad with two states, exposed or not exposed, and its "importance" is the effect on conversion). Additionally the idea of "importance" is in and of itself a vague term, so variable importance is often guided by principles that lend context to the meaning of importance. This concept was presented early on by Achen [1], noting that a key question on variable importance is "important for what?" He suggests that any inquiry into the importance of a variable should also have an explicit objective function associated with the importance. In the case of attribution, advertising channel (or variable) importance should be determined by a channel's ability to influence conversion.

Much of the literature found on the topic is focused on variable importance in linear multiple regression scenarios. Many of the methods for measuring variable importance in the regression context involve direct interpretation of regression coefficients or deriving a measure off of some transformation of the regression coefficients. A thorough survey and analysis of these methods is available [14]. It is beyond the scope of this paper to fully define and explain each method surveyed, but a common criticism is that the coefficient-based methods lead to problems of interpretability in the presence of collinearity amongst predictors [14, 3, 16]. Shao et al. [20] present a coefficient based method for multi-touch attribution that has several limitations, similar to those discussed in the variable importance literature. These are: (1) logistic regression coefficients are difficult to interpret intuitively, and (2) negative coefficients may arise due to channel collinearity. These two problems suggest that coefficient-based attribution methods do not satisfy the four properties presented in section 2.1.

Several of the more recent advances in variable importance for linear regression involve game theory or variable transformation approaches to ensure independence of predictors. Budescu [3] developed a method called Dominance Analysis, which aims to measure predictor importance in a way that is invariant to subset selection and stands up to collinearity of predictors. An important contribution of this method was to incorporate the "partial effects" of a predictor, which is defined as the average increase in R^2 by the inclusion of a predictor, conditioned on all subsets of predictors. This notion of partial effects is substantively the same as the method of Shapely Value regression, later used by Grömping and Lipovetsky et al. [16, 9]. The Shapely Value regression attempts to decompose the R^2 of a model such that each predictor's Shapely Value represents the average increase in R^2 the predictor induces upon inclusion into the model across all possible subsets of predictors. The Shapely Value approach lends itself to channel importance and attribution analysis, and will be developed further below.

The variable importance concepts surveyed so far generally involve decompositions of explained variance in linear regression models. Other variable importance measures that can be applied under more general modeling assumptions have been presented by Breiman [2] and Van der Laan [25]. Strobl et al. [23] present an overview of variable importance measures that can be derived from Random Forests as developed by Breiman [2]. Noted in Strobl et al. [23] is the fact that interpretability of Random Forest variable importance is not as straightforward as in the single decision trees that compose the forest. This lack of interpretability led us to exclude Random Forests as a possible channel importance measure. Van der Laan et al. [25] introduce a measure of variable importance that is of direct relevance to channel importance and attribution measurement. They define a parameter using the same counterfactual framework discussed above. When data conditions satisfy the strict causal parameters defined above, this variable importance parameter is equivalent to a causal effect parameter. Given any violation of the assumptions, however, the counterfactual parameter can be interpreted as a variable importance measure. An advantage this method has over many of the regression based methods, in the context of online ad attribution, is that the parameter of interest (channel importance) is directly measured using a counterfactual framework. This is a main reason we have chosen this method for our design, in addition to its flexibility of interpretation in different settings.

3. MULTI-TOUCH ATTRIBUTION

In this section, we present a multi-touch attribution methodology that satisfies all four properties presented above. Within the counterfactual framework, we present a method that under varying assumptions about the data can be interpreted as either casual effects attribution or channel importance attribution. No existing, published method for multi-touch attribution satisfies the properties that we proposed in section 2.1. Commercialized methods such as [5, 11, 12], as well as last-touch attribution, are driven by global heuristics rather than individual campaign-level data and do not explicitly model fairness into the allocation of conversion. To our knowledge, Shao et al. [20] have published the only data-driven methodology for multi-touch attribution that is used in a production system. The "simple probability model" they present is data driven and fairly intuitive, though it does suffer from an arbitrary normalization to ensure economic efficiency. Despite this fact, we show that this method gives equivalent results after rescaling to our method under specific simplifying assumptions (discussed in section 3.3).

3.1 "Causal" Attribution

For each subject i (generally an internet user), we define an ordered data structure $\Lambda_i = \{W_i, C_i = \{C_{i1}, \dots, C_{iK}\}, Y_i\}$. W_i is a vector of user attributes that represents the state of the user prior to viewing any ads (such as demographics, prior purchase behaviour, prior searches, etc.). C_i is an ordered set of K advertising channels that have shown an ad to the user (with K possibly being different for each i).[3] Each C_{ik} is a categorical random variable that takes on the value of the channel that displayed the advertisement (e.g., Search Engine 1, Search Engine 2, Display Ad Company 1, Display Ad Company 2, etc.). Finally, Y_i is a binary indicator of the conversion event of interest following the set of ads (like a purchase on an advertiser's website).

Attribution within the context of online advertising involves two major allocations. First, how much value do we credit to the advertising campaign in total, and second, within the value credited to the campaign, how do we allocate credit to the individual channels serving ads to user i within the campaign. We address the first allocation by defining a parameter that represents the total value created by the set of channels C_i serving ads to a user.

$$\begin{aligned}\Psi_i &= E[Y|C_i = \{C_{i1}, \dots, C_{iK}\}, W = w_i] \\ &- E[Y|C_i = \{\emptyset\}, W = w_i] \qquad (1)\end{aligned}$$

Ψ_i can be interpreted as the w-adjusted impact of the total set of advertisements on the user's expected value of conversion. For the second allocation, we seek a decomposition of Ψ_i that meets the criteria of: (1) being based on counterfactual analysis, (2) derived from the campaign data, (3) sums to Ψ_i without arbitary normalizations and (4) is generally interpretable.

We seek to define a quantity V_k that satisfies the following constraint: $\sum_{k=1}^{K} V_k = \Psi$, where k indexes some channel $C_k \in C$ (for simplicity of notation, we will drop the index on user i). V_k is the amount of value that should be allocated to channel C_k and should reflect its expected impact on Y given the other channels in C. We start by defining another causal parameter which represents channel C_k's expected influence on Y after serving an ad, given another subset of other channels in C having also served an ad before it. We first define S as some subset of C/k, where C/k is the set C excluding C_k and including the null set $\{\emptyset\}$. Then channel C_k's marginal contribution over S to the expected value of conversion is:

$$\Phi_S^k = E[Y|S \cup C_k, W = w_i] - E[Y|S, W = w_i] \qquad (2)$$

We now propose that channel C_k's total attribution allocation should reflect its expected impact on conversion across all possible subsets S of C/k (i.e., start by measuring how channel C_k increased the probability of conversion, given the $k-1$ channels serving ads before it, for one possible ordering of C, then take an expected value over all possible orderings of C). To acheive this, we represent V_k as the expected value of channel C_k's impact on conversion given the ads shown before C_k. More formally:

$$V_k = E[\Phi_S^k] = \sum_{S \subseteq C/k} \omega_{S,k} * \Phi_S^k \qquad (3)$$

where $\omega_{S,k}$ is the probability that set C begins with the sequence $\{S, C_k, \dots\}$ in some distribution Ω of possible orderings of channel set C, and S indexes over all possible subsets of C/k.

We can illustrate our attribution system using a toy example of a two channel system. Given $C = \{C_1, C_2\}$:

$$\begin{aligned}V_1 &= \omega_{2,1} * [E[Y|\{C_1, C_2\}, W = w_i] - E[Y|\{C_2\}, W = w_i]] \\ &+ \omega_{\emptyset,1,2} * [E[Y|\{C_1\}, W = w_i] - E[Y|\{\emptyset\}, W = w_i]] \\ V_2 &= \omega_{1,2} * [E[Y|\{C_1, C_2\}, W = w_i] - E[Y|\{C_1\}, W = w_i]] \\ &+ \omega_{\emptyset,2,1} * [E[Y|\{C_2\}, W = w_i] - E[Y|\{\emptyset\}, W = w_i]]\end{aligned}$$

Provided that $\omega_{S,k} = P(\{S, C_k, C/\{S, C_k\})$ (i.e., the probability that C is ordered by the sequence $\{S, C_k, C/\{S, C_k\}\}$), then it can be shown that $\sum_{k=1}^{K} V_k = \Psi$ (we omit the proof due to space limitations). This holds for all sets C of arbitrary size K.

Intuitively, V_k is the expectation of channel C_k's influence on Y over all of the possible orderings of C in a set of ads. For a single user in an observed advertising campaign, the ordering of C is realized and not random, so if Ω is chosen as the empirical distribution of channel orderings, V_k can be defined without the expected value. Specifically, $V_k = \Phi^k_{\{k-1\}}$, where we define set $S = \{k-1\}$ as the set of channels showing ads before channel C_k. While convenient and intuitive, using the empirical distribution of Ω introduces problems that may not be in the best interests of the advertiser in question. We can illustrate this with a simple two channel system.

Given $C = \{C_1, C_2\}$, let's assume the following: $E[Y|\{\emptyset\}] = E[Y|\{C1\}] = E[Y|\{C2\}] = 0$, and $E[Y|\{C1, C2\}] = \delta \gg 0$. The attribution for channels C_1 and C_2 are then:

$$\begin{aligned}V_1 &= E[Y|\{C1\}] - E[Y|\{\emptyset\}] = 0 \\ V_2 &= E[Y|\{C1, C2\}] - E[Y|\{C1\}] = \delta\end{aligned}$$

What we have defined here is an ad campaign in which channels C_1 and C_2 have zero individual impact but have a strong interaction effect on Y. In a system where C_2 systematically always serves an ad following C_1, we end up with C_2 accepting full credit for the value of Y in all instances (this is a likely scenario when paid search channels are involved). While this may be fair from C_2's perspective, as no value would have been created without its participation, in the long run, this is a suboptimal allocation for both channels as well as the advertiser. The reason is that C_1 would likely not participate in subsequent campaigns (or would be cut by the advertiser) due to its lack of payment, and following this exit from future campaigns, C_2 would not be able to create value. So under this allocation scheme, both channels and the advertiser lose out from any potential value creation. While this is an extreme example meant for illustrative purposes, it introduces another concept of what it means to be "fair" in a value allocation sense.

For any arbitrary distribution Ω of possible orderings of channel set C, the general attribution methodology defined by equation (3) favors channels that systematically appear

[3] A natural and logical ordering would be the temporal ordering in which ads were shown.

later in the (expected) orderings of C in terms of allocating the value of positive multi-channel interaction effects (the reverse is true for negative effects). This introduces an undesirable quality in that a channel might attempt to game the system by targeting a specific ordering (first or last, depending on the interaction effects). Gaming the system distracts the channel from what it should really be doing, which is driving conversions on behalf of the advertiser. The implication of this on the long-term equilibrium of value allocation is beyond the scope of this paper, but it introduces a need to reexamine the allocation of value for multi-channel interaction effects, such that it reduces opportunities to game the system.

"Fairness" so far has been defined explicity using the counterfactual framework on a channel's ad treatment. We showed that this definition as is introduces possibilities for channels to game the system (though at a much weaker rate than current attribution methods allow). In order to reduce this possibility, and constrain "fairness" such that it aligns advertiser and channel incentives, we add that a "fair" attribution system should evenly distribute multi-channel interaction effects across the channels creating the effect. We can acheive this in equation (3) by defining Ω as a uniform distribution over all possible orderings of C. This new definition of Ω creates a uniform allocation of multi-channel interaction effects by giving equal probability to each channel following the other channels in question, so that, in expectation, each gets an equal proportion of the interaction effect. Under this uniform distribution, $\omega_{S,k}$ is solely a function of the cardinality of S and C. Specifically:

$$\omega_{S,k} = \frac{|S|!(|C| - |S| - 1)!}{|C|!} \tag{4}$$

With $\omega_{S,k}$ defined by equation (4), our attribution methodology in equation (3) is equivalent to the Shapely Value solution to value distribution in Cooperative Game Theory [21, 17]. We refer interested readers to [21, 17] for a thorough treatment on the subject. This is to our knowledge, the first application of the Shapely Value distribution methodology being applied to value allocation in advertising attribution.

3.2 "Channel Importance" Attribution

Thus far we have defined our attribution system using w-adjusted expectations of channel impacts on conversion in order to treat attribution as a causal problem. In realistic production settings however, it is highly likely that at least one of the strict causal assumptions will be violated, and that often, confounding attributes are not even observed. We thus seek a formulation of our general attribution model that can be computed without adjusting for possible confounders. Such a model may satisfy our four properties of a good attribution system, though its interpretation as a causal estimator may no longer be valid. Nonetheless, we can recast this as a channel importance measure, whose formulation is motivated by the corpus of variable importance measures reviewed.

Another important aspect to consider is the computational feasibility of our general model. Note that in equation (3) we sum over all possible subsets of C/k. For large $|C|$, the complexity of our model increases exponentially. We thus redefine a formulation of our attribution methodology such that computational complexity may be reduced.

We define an aggregated data set $\lambda^g = \{C = \{C_1, \ldots, C_K\}, \gamma = \sum Y, n\}$ as a group of n users having an identical realization of C (i.e., saw ads by the same set of channels), producing $\gamma = \sum Y$ total conversions (it is convenient to think of λ^g as a grouping of Λ, such as that returned by a query "select C,$\sum y$,$\sum 1$ from Λ group by C").

We now want to allocate the total conversions γ_i across the set of channels C having shown an ad in realization λ^g_i. We can do this with the following model:

$$SV_k = n \sum_{S \subseteq C/k} \omega_{S,k} * [P(Y|S \cup \{C_k\}) - P(Y|S)] \tag{5}$$

where $\omega_{S,k}$ is the same as defined in equation (4). We change our attribution notation from V_k to SV_k to specifically acknowledge the equivalence to the Shapely Value formulation (i.e., definining $\omega_{S,k}$ by equation (4)). We also change our notation from w-adjusted model on an individual user to a non-w-adjusted model on a group of users (although, given observable features in W, one may reintroduce the notation). Without adjusting for w, we are implicitly making the assumption that Ψ is the total number of conversions observed.

The grouping of users enables us to redefine value as a countable number of conversions, as opposed to an expected probability of conversion. In a business setting involving people of mixed mathematical backgrounds, this aids in the interpretability of the methodology. Further, by grouping the users into cohorts having the same realization of C, we reduce the complexity of both computation and data storage. The number of rows in λ^g is bounded by total channel count. If R is the count of rows in λ^g, then $R \leq 2^{|C_{union}|} - 1$, where C_{union} is the union of all realizations of C (simply put, the total number of channels being evaluated). This has tremendous practical value when a typical campaign can serve ads to more than 100MM users but have only 15 channels. Choosing the Shapely Value formulation of equation (3) on grouped data also means that estimation of $\omega_{S,k}$ is no longer needed, as $\omega_{S,k}$ is solely defined by $|S|$ and $|C|$.

3.3 Generalization of prior state of the art

We noted that Shao et al. [20] have published the only statistical and data-driven attribution solution used in production as known by the authors. We can easily show that their "simple probabilistic model" is, after rescaling, equivalent to the Shapely Value formulation of equation (3) under certain simplifying assumptions.

The definition of the data is the same as in 3.1, but we drop W from consideration and for simplicity, omit the index over user i.

Shao et al. [20] define S_k as channel C_k's unnormalized share of the total conversions, and calculate S_k with:

$$\begin{aligned} S_k &= P(Y|C_k) \\ &+ \frac{1}{2(K-1)} \sum_{j \neq k} [P(Y|C_j \cup C_k) - P(Y|C_j) - P(Y|C_k)] \end{aligned}$$

And then the attribution is calculated by normalizing S_k with the following: $V_k^T = \frac{S_k}{\sum_k^K S_k}$

If we assume that the marginal effects of any ad treatment after the 2nd are negligible, then S_k is a rescaled version of SV_k from equation (5) (with $n = 1$). With S_k and SV_k only differing by a scalar factor, the percentage of value each method allocates to the given channels is identical. To show this, we define our above assumption more formally: for all such cases where $|S| > 2$, we get $[P(Y|S \cup \{C_k\}) - P(y|S)] = 0$. This assumption states that the additional ad exposures (after exposures from S) do not increase the probability of a conversion.

With the stated assumptions above, our attribution SV_k reduces to:

$$\begin{aligned} SV_k &= \frac{2}{K} P(Y|C_k) \\ &+ \frac{1}{K(K-1)} \sum_{j \neq k} [P(Y|C_j \cup C_k) - P(Y|C_j) - P(Y|C_k)] \\ &= \theta S_k \end{aligned}$$

where $\theta = (2/K)$.

The difference in interpretation between these two attribution methods is that SV_k is an estimate of the additive portion channel C_k contributes to the total probability of a conversion, whereas S_k is an unnormalized estimate of the share of the conversion probability to be allocated to channel C_k. With this connection between SV_k and S_k, we can assert that, should the assumptions hold, the method presented by Shao et al. [20] satisfies the fairness, data driven and interpretability properties we proposed in section 2.

4. EMPIRICAL RESULTS

In this section we present attribution results on 3 different datasets. The first dataset is from 2 weeks of an actual advertiser's online ad campaign, where the advertiser executed a direct buy across multiple well-known internet domains. The second is a synthetic, simulated ad campaign dataset, generated to highlight certain properties of our attribution system. The last dataset is from several ad campaigns internal to M6D[4] and is distinct from the other two in that it has both ad channel data and user characteristics that are well known confounders in digital advertising. All attribution results labeled "Multi-Touch" were computed using the methodology presented in section 3.2.

4.1 Model Selection and Estimation

For estimation, any model that produces estimates of binary class membership probability would be appropriate as a candidate model. The candidate models explored in this paper were: (1) normal maximum-likelihood logistic regression, (2) elastic-net regularized logistic regression, and 3) smoothed empirical probability estimation. In each of these model classes we vary complexity, variable selection and smoothing parameters where appropriate, and use likelihood based k-fold cross validation for model selection. The choice of likelihood based cross validation over other emprical risk measures is motivated by the results of Van der Laan et al. [26] who show that, in finite samples, density estimators chosen via likelihood k-fold cross validation converge asymptotically to the true data generating distribution (using Kullback-Leibler distance to measure convergence between empirical and true distributions). Our ultimate goal is to estimate unbiased conditional probability distributions, so the asymptotic optimality of likelihood based cross validation makes it an appropriate choice of risk function for model selection.

4.2 Direct Buy Ad Campaign

The first results we present are based on data from an actual campaign run by a major media company that sells online subscriptions. The campaign was run as a direct buy on 7 major online content providers (with "Direct Buy" meaning that the advertiser contracted directly with the 7 channels to run the campaign, with no intermediary ad exchanges being used). There was no targeting done based on user characteristics, so the ads were displayed solely based on visitation to one of the 7 content providers. Figure 1 shows how the results of our multi-touch attribution methodology established in this paper compares to a last-touch attribution methodology. The bars show the relative difference in conversion credit between last-touch and multi-touch attribution (where a negative difference means multi-touch assigns less credit than last-touch). The channels are sorted from left to right in order of increasing channel effectiveness (as measured by the probability of the user converting after having seen an ad on that channel). By sorting on increasing channel effectiveness, we can see that the two channels that do better with the last-touch methodology are those with the smallest channel effectiveness.

One might expect the propensity to be the last-touch channel to affect attribution differences, but for this particular campaign, the range of different propensities to be last-touch was only 41.4%–51.4%. The untargeted nature of the campaign leads to an almost even likelihood of each channel to be the last-touch, so the last-touch bias isn't as much a factor in explaining the differences shown. In many other targeting situations though, certain strategies target users that are further down the conversion path [10]. For instance, general display advertising tends to broadly target users that may have no intention on purchasing a given product, whereas Retargeting and Paid Search tend to target users during their product research phase. This causes Retargeting and Paid Search to have a much higher tendency to be last when overlap occurs. In such cases, we would expect the multi-touch attribution system to differ even more from last-touch attribution.

Given any comparison of conversion allocations from two different attribution methods, the goal of the analyst is to determine which is better and to make recommendations based on this choice. A challenge for attribution analysis though is that the "truth" is a quantity that may never be known, so there is no benchmark or holdout set to evaluate "better" in a quantitative sense. This motivated our proposition of governing principles for mult-touch attribution analysis. The absolute "truth" in the above campaign may be never known, but with respect to the properties given in section 2.1, we would argue that in Figure 1 the multi-touch method is a better allocation than the last-touch method.

In the next section we make an attempt to compare our methodology with the industry standard of last-touch attribution where the "truth" is known. We do this by synthesizing a campaign data set and explicitly control the factors that influence attribution allocations.

4.3 Simulation Study

Our goal in this simulation analysis was to simulate the effects of various targeting and ad serving strategies working simultaneously on the system and observe trends in attribution measurements given the known data generating distri-

[4] M6D is a display advertising firm that uses proprietary machine learning systems to target ads on behalf of branded advertisers.

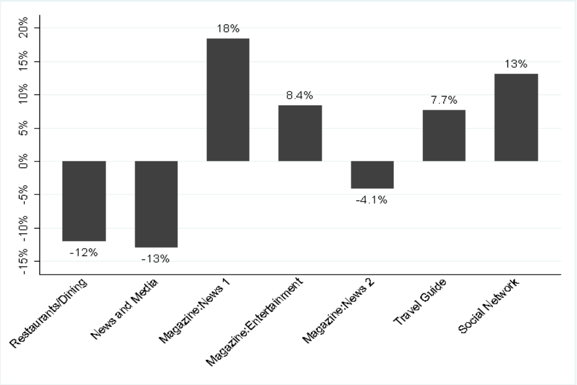

Figure 1: Multi-touch vs. last-touch attribution on a direct buy ad campaign. The bars show the relative difference between the two attribution systems differ. Channels with lower ability to drive conversions and with higher likelihood to show ads last generally receive less credit in the multi-touch system as compared to the last-touch system.

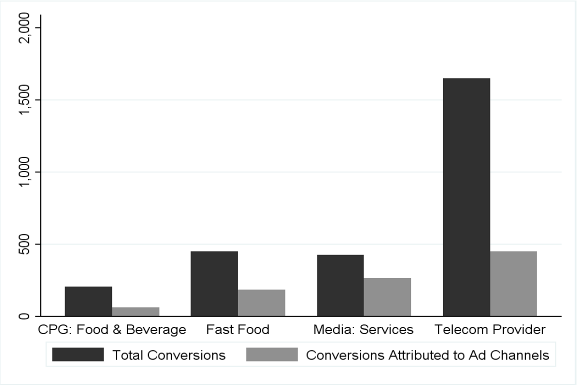

Figure 2: Attribution analysis in the presence of the confounding feature of prior interactions with a brand's website. Each bar shows the total conversions attributed to the campaign, with and without the confounding feature. We see that in all of campaigns, a large portion of the conversions can be attributed to the user having prior interactions with the advertiser's website.

bution. The online advertising strategies we simulated were: (1) General Prospecting, (2) Retargeting, which is the practice of targeting users that had previously been observed visiting or purchasing from an advertiser's website and (3) Paid Search, which is the practice of serving sponsored ads in search query results. We simulate these strategies by controlling three factors: (1) the channel correlation (or likelihood for simulated channels to show ads to the same user), (2) the likelihood to serve an ad and (3) the probability of conversion conditioned on seeing an ad by a given channel. For the purpose of comparison to a last-touch attribution method, we also generate an ordering of ads for each user and we introduce a bias so that each synthetic channel has a different propensity to be last.

We start the simulation by generating an NxK matrix of multivariate normal random vectors Θ, where each of the N rows simulates a user and the K column variables are normally distributed ($X \sim \mathcal{N}(\mu, \Sigma)$),with $\mu = [0, \ldots, 0]$ and Σ defined as follows:

$$\Sigma_{jk} = \begin{cases} 1 & \text{if } j = k \\ .4 & \text{if group(j) = group(k)} \\ .1 & \text{if group(j)} \neq \text{group(k)} \end{cases}$$

We define a group as a set of channels belonging to a single strategy. Channels within the same group have a covariance of .4 and channels in different groups have a covariance of .1. This covariance structure is designed to achieve the property that similar strategies tend to hit the same users while different strategies tend to hit different users. Thus, if columns K_i and K_j are correlated in our data matrix, the channels C_i and C_j will be correlated and will serve more simulated ads to the same users.

Our matrix of simulated user characteristics Θ has 50MM rows and 14 columns. We transform this into a binary ad matrix by comparing each element $\Theta_{i,k}$ to our chosen ad propensity likelihood (column 3 in Table 1) and generating an ad if $NormInv(\Theta_{i,k}, 0, 1) <$ AdPropensity C_k. We then generate a conversion using the following formula:

$$P(Y) = [1 - \prod_{k}^{K}(1 - P(Y|C_k))] * \delta^{\sum I(C_k)} \qquad (6)$$

The first term in brackets represents the probability of conversion assuming zero interaction effects. In many advertising scenarios however, one should not assume that each subsequent ad will have an effect proportional to its lone effect, so we use the right-hand term to account for the marginally decreasing effect of each ad. For this simulation study we choose $\delta = .95$. The fourth column of table 4 shows the values for $P(Y|C_j)$ that we set. The last item we simulated was a propensity for each channel to be last in the ordering of channels that served an impression to a single user. We accomplished this by generating a random number for each channel on a given user and sorting by this number. Each random number was generated such that each channel has a different propensity to be last. The propensity is shown in the 5th column of table 1.

Table 1 shows the results of our simulation study. We report the total number of conversions attributed by a last-touch system and by the multi-touch methodology from section 3.2. The "Delta" columns show absolute and relative differences between the two, where a positive delta means the multi-touch model attributed more conversions than last-touch. Within each group, we see that the total number of conversions attributed by the last-touch model is influenced more by ad propensity likelihood and last-touch propensity than by $P(Y|C_k)$. We can see though that these also have the largest negative delta. The multi-touch method generally rewards the channels with the highest $P(Y|C_k)$ and also corrects for the last-touch bias.

These results are in line with probably the most important of the four properties from section 2, i.e., that a channel should be rewarded proportionally to its ability to affect the likelihood of conversion in a browser. As previously discussed, the channels labeled 7 and 8 are rewarded more for their ordering and volume than for their effectiveness in

		Data Generating Parameters			Attribution Results			
Channel	Group	Ad Propensity Likelihood	Simulated Conversion Rate	Last Touch Propensity	Last Touch Conversions	Multi Touch Conversions	Delta N	Delta %
1	Gen Prospecting	5.0%	0.100%	0.2%	1,023	2,176	1,153	113%
2	Gen Prospecting	10.0%	0.080%	0.2%	1,932	3,284	1,352	70%
3	Gen Prospecting	10.0%	0.070%	0.2%	1,854	3,085	1,231	66%
4	Gen Prospecting	15.0%	0.050%	0.2%	2,491	3,434	943	38%
5	Gen Prospecting	15.0%	0.050%	1.8%	3,134	3,143	9	0%
6	Gen Prospecting	20.0%	0.010%	1.7%	2,998	736	-2,262	-75%
7	Gen Prospecting	20.0%	0.008%	6.7%	3,558	260	-3,298	-93%
8	Gen Prospecting	25.0%	0.008%	6.8%	4,406	4,09	-3,997	-91%
9	Retargeting	2.5%	0.500%	3.0%	3,921	5,673	1,752	45%
10	Retargeting	2.5%	0.400%	6.0%	3,375	4,489	1,114	33%
11	Retargeting	3.0%	0.300%	10.5%	3,468	4,068	600	17%
12	Retargeting	3.5%	0.250%	15.3%	3,728	3,997	269	7%
13	Search	0.5%	1.000%	23.7%	2,109	2,430	321	15%
14	Search	0.5%	2.000%	23.6%	5,329	5,045	-284	-5%

Table 1: Table of simulation parameters and attribution results on the simulated campaign dataset. The first two data columns represent parameters used to simulate the campaign data, the column "Last-Touch Propensity" indicates the likelihood we give the channel to be the last serving an ad, and the last four represent results of the attribution analysis. We can see that the delta between methodologies is driven by the ordering bias and $P(Y|C_k)$.

the last-touch system. In the online advertising industry, these two channels might be labeled "Carpet Bombers" due to their tendency to serve many ads with relatively low effectiveness. This type of strategy is generally misaligned with the advertiser's goals, and can result in a misallocation of resources when the advertiser rewards these Carpet Bombers at the expense of channels actually driving conversion.

4.4 Attribution in the presence of user confounding

The last set of data we analyzed was from advertising campaigns run by M6D on behalf of four of its advertising clients. In the prior analyses, we performed attribution in the absence of any measured user-level confounding. However, as is often is the case with online advertising, various strategies are employed to target specific segments of the online population, and these segments often vary in their organic propensity to convert. Prior studies on estimating the causal effects of display advertising present detailed treatments of the various types of user confounding often observed [22, 15, 4]. As was specifically discussed by Stitelman et al. [22], a type of user confounding that is often present in online display advertising is the past interactions between a user and the advertising client's website. The strategy that specifically targets users with observed prior interactions is called Retargeting, and this analysis contains data from four campaigns that employed the strategies of Retargeting and display targeting as presented by [18]. The specific confounder that we analyze in this section is a binary indicator of whether or not the user had previously visited and/or purchased from the clients site. In this analysis, we deviate from prior studies of causal estimation in online advertising in that we are attempting to estimate attribution allocations of conversion credit among a set of converters that received treatment, as opposed to measuring the effect of treatments on conversion on a general segment of users.

To derive the attribution allocation for the confounding property of having prior interactions with the advertising client's website, we treat the confounding variable (w) as if it were another advertising channel. The attribution on w can be interpreted as the expected marginal effect of w across all users that were treated with ads. More formally:

$$Attribution(w) = E[E[Y|A, w = 1] - E[Y|A, w = 0]] \quad (7)$$

As in the previous sections, we use our proposed Shapely Value attribution formula to estimate the attribution allocation to the confounding user segment of past client customers.

Figure 2 shows the result of our retargeting confounder attribution analysis on four campaigns run by M6D. In these four campaigns, anywhere between 38% and 73% of conversions can be attributed to the confounding presence of the user having had prior interactions with the advertising client's website. The implications of these results are that, particularly in the presence of retargeting, advertisers are often assigning credit to conversions that are driven more by the user's own volition to convert rather than the influence of advertising. Our proposed attribution methodology may enable advertisers to better allocate dollars to advertising strategies that are effectively influencing internet users to convert, as opposed to just capturing credit for the users' present and uninfluenced propensity to convert.

5. CONCLUSION AND FURTHER WORK

The methods for multi-touch attribution presented in this paper are motivated by a need to bring about more standardization and data-driven intelligence to the measurement of online advertising campaigns. Better multi-touch attribution measurement is an often-cited top priority within the industry. Solutions exist that offer firms alternatives from the last-touch attribution model, though the variety of these methods suggests a present lack of standardization within the industry. Performance measurement is the economic scorecard by which firms succeed or fail, so transparency and standardization are needed. We presented the four properties of section 2 to meet this need. We developed the generalized attribution solution in equation (3) as a synthesis of methods found in both the causal estimation and variable importance fields. Key to our attribution solution is the notion of the counterfactual analysis, and the methodology presented is to the authors' knowledge the only attribution solution that is causally motivated.

The attribution methodology presented herein is designed to be a complete solution to the problem of multi-touch attribution. Nonetheless, we make assumptions about the data and we propose for future work a sensitivity analysis on how varying these assumptions impacts final attribution allocations. Specifically, as we encode ad treatments by channel as opposed to by ad, we do not explicitly account for varying average ad frequencies by channel (though implicitly, frequency should be captured by $P(Y|C_k)$). We also allow for subjectivity on the distribution of channel ordering Ω. While this subjectivity is motivated by different interpretations of "fairness", an area for future work would be to analyze the sensitivity of attribution allocations to varying distributions of Ω. This line of research can be extended to a game theoretic context, where long term equilibria are explored based on the strategic implications of Ω. Finally, in our estimation and analysis, we did not attempt to make any strict causal interpretations on the results. Strict causal estimation requires a more complex set of methods that are beyond the scope of this paper. We presented this work with the practitioner in mind, knowing that strict causal methods may be untenable due to the scale and nature of the data, as well as the complexity of the causal estimation methods. However, extending our attribution methodology to a strict causal estimation problem would be a natural sequel to the work presented.

6. REFERENCES

[1] C. Achen. *Interpreting and using regression.* Number 29. Sage Publications, Inc, 1982.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] D. Budescu. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542, 1993.

[4] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. 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*, pages 7–16. ACM, 2010.

[5] J. A. I. Chandler-Pepelnjak. Measuring roi beyond the last ad. `http://www.atlassolutions.com/uploadedFiles/Atlas/Atlas_Institute/Published_Content/dmi-MeasuringROIBeyondLastAd.pdf`.

[6] eMarkerter. Tbd. `http://www.emarketer.com/Report.aspx?code=emarketer_2000774`.

[7] eMarketer. Us online ad spend to close in on $40 billion. `http://www.emarketer.com/Article.aspx?R=1008783`, Jan. 2012.

[8] I. Finlo. Removing the barriers to growing online media spend: Transparency. `http://www.admonsters.com/blog/removingbarriersgrowingonlinemediaspendtransparency`, Nov. 2010.

[9] U. Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

[10] A. Hunter, M. Jacobsen, R. Talens, and T. Winders. When money moves to digital, where should it go? `http://www.comscore.com/Press_Events/Presentations_Whitepapers/2010/When_Money_Moves_to_Digital_Where_Should_It_Go`, Sept. 2010.

[11] C. Inc. Media attribution. `http://www.clearsaleing.com/product/media-attribution/`.

[12] C. M. Inc. What is c3 metrics? `http://c3metrics.com/executive-summary/`.

[13] E. Inc. `http://www.encoremetrics.com/solution`.

[14] J. Johnson and J. LeBreton. History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3):238, 2004.

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

[16] S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[17] M. Osborne and A. Rubinstein. *A course in game theory.* The MIT press, 1994.

[18] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 707–716. ACM, 2009.

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

[20] X. Shao and L. Li. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 258–264. ACM, 2011.

[21] L. Shapley. A value for n-person games. *The Shapley value*, pages 31–40, 1953.

[22] O. Stitelman, B. Dalessandro, C. Perlich, and F. Provost. Estimating the effect of online display advertising on browser conversion. *Data Mining and Audience Intelligence for Advertising (ADKDD 2011)*, page 8, 2011.

[23] C. Strobl, A. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[24] A. Tsiatis. *Semiparametric theory and missing data.* Springer Verlag, 2006.

[25] M. Van Der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):2, 2006.

[26] M. Van Der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):4, 2004.