Shapley Meets Uniform: An Axiomatic Framework for Attribution in Online Advertising

Raghav Singal
Columbia University
rs3566@columbia.edu

Omar Besbes
Columbia University
ob2105@gsb.columbia.edu

Antoine Desir
INSEAD
antoine.desir@insead.edu

Vineet Goyal
Columbia University
vg2277@columbia.edu

Garud Iyengar
Columbia University
garud@ieor.columbia.edu

ABSTRACT

One of the central challenges in online advertising is attribution, namely, assessing the contribution of individual advertiser actions including emails, display ads and search ads to eventual conversion. Several heuristics are used for attribution in practice; however, there is no formal justification for them and many of these fail even in simple canonical settings. The main contribution in this work is to develop an axiomatic framework for attribution in online advertising. In particular, we consider a Markovian model for the user journey through the conversion funnel, in which ad actions may have disparate impacts at different stages. We propose a novel attribution metric, that we refer to as *counterfactual adjusted Shapley value*, which inherits the desirable properties of the traditional Shapley value. Furthermore, we establish that this metric coincides with an adjusted "unique-uniform" attribution scheme. This scheme is efficiently computable and implementable and can be interpreted as a correction to the commonly used uniform attribution scheme.

CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Mathematics of computing** → *Markov processes*; • **Applied computing** → *Electronic commerce*; *Operations research.*

KEYWORDS

Digital economy, online advertising, attribution, Markov chain, Shapley value, causality

ACM Reference Format:

Raghav Singal, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar. 2019. Shapley Meets Uniform: An Axiomatic Framework for Attribution in Online Advertising. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313731

1 INTRODUCTION

With the rise of the Internet, the digital economy has become a trillion dollar industry accounting for over 6% of the U.S. GDP [19]. A multitude of retailers reach consumers through online sources with the goal of acquiring new customers and managing relationships with the existing ones. Such wide range of economic activities taking place in the digital world has enabled data collection at a massive scale, allowing retailers to better understand customer behavior and enhance service quality using data-driven decisions. With over a billion people with access to the Internet, it is not surprising that advertising has moved to the digital space. The global digital marketing sector witnessed a growth of 21% in 2017, which increased its market size to USD 88 billion [44]. Key decisions in this space include allocating budget to various ad channels and media (e-mail, display media platforms, and paid search for instance) and bidding for ads to push one's product towards the desired customer demographic at the right time. Such problems require the retailer to understand the value of showing an ad via a specific channel at a given time. How can an advertiser assign or *attribute* value to the advertising actions taken across the different channels and media? *Attribution* is one of the central questions in online advertising today. The value of each channel is an important input to media mix optimization, helps build an understanding of the customer journey, and also helps a company to justify its marketing spend [38, 45]. Incorrect understanding of the effectiveness of online channels can result in sub-optimal budget allocation, possibly resulting in lost revenue. Recognizing the importance of this problem, the Marketing Science Institute identifies attribution as the topmost research priority for the period 2016-2018 [23]. Though there exist multiple articles in this domain, there lacks a systematic approach that is both theoretically-sound and tractable.

Attribution is inextricably linked to causality since it involves quantifying the added value of showing an ad over the baseline value of what *would have* happened if no ad was shown (*counterfactual*). In this work, we take the following view of causality, which comes from the pioneering work of Rubin [41]:

> *"Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1."*

In the context of digital advertising, treatments E and C are the advertiser showing, and not showing an ad, respectively. Accordingly, attribution involves capturing the causal effect of an ad where the baseline corresponds to not showing an ad.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313731

Causality is an active research area, and there exist paradigms for capturing the causal effect of a treatment. See, e.g. Chockler and Halpern [11], Collins et al. [13], Eells [16], Halpern and Pearl [20], Hitchcock [21], Hume [22], Morgan and Winship [35], Pearl [37] and Rubin [41]. We comment on some of these alternative approaches in our concluding remarks.

As with the existing work in attribution, we also consider the setting where an advertiser is interested in understanding the contributions of various ads to a single product they are promoting. Even with a single product, attribution is a challenging problem. It involves distributing the value generated by a network of actions to each individual action. Such a network might have a mix of interaction effects that one needs to account for when decomposing the network value. Capturing such (possibly non-linear) interaction effects is quite non-trivial.

1.1 Related literature

There exists a large and diverse literature on the attribution problem in online advertising. We present a brief overview of existing approaches and refer the reader to existing surveys [12, 28] for a more complete treatment. Attribution methodologies can be classified into two broad classes: *rule-based* or *algorithmic*. We discuss both the classifications below.

1.1.1 ***Rule-based heuristics.*** Rule-based heuristics include approaches such as *last touch attribution* (LTA), *uniform weights*, and *customized weights* [6, 38–40]. In LTA, an advertiser allocates all the value generated by a user to the last ad directed at her whereas under a uniform weights scheme, all the ads are allocated an equal credit. Ads receive tailored weights under a custom weights scheme. Although such heuristics are transparent and tractable, there is no rationale justifying their appropriateness as a measure for attribution. Metrics such as LTA appear unfair since they do not value the contribution of channels that build product awareness. Uniform or customized weights might appear to be a fix but there is no a priori reason to believe that attribution should be linear.

1.1.2 ***Algorithmic approaches.*** The algorithmic approaches can be classified as using either *incremental value heuristic (IVH)* (or *removal effect*) or *Shapley value (SV)* as a measure for attribution.

IVH. IVH computes the change in the eventual *conversion*[1] probability of a user when a specific ad is removed from her path. This is the most common metric for attribution [2–5, 15, 27, 29, 30]. In this approach, one calibrates a model that predicts the conversion probability as a function of the ad exposure and then, uses the estimated model to compute the incremental value of each ad. The novelty comes from the model proposed to describe user behavior, e.g., an HMM [2] or a neural network [4]. Irrespective of the prediction model's level of complexity, attribution is measured via IVH. There exists little (if any) formal justification for why IVH serves as a good approximation to attribution. In fact, in Section 3.2, we show that IVH can result in inappropriate allocations.

SV. SV [43] is a well-accepted concept for assigning credit to individual players in a cooperative game. The value generated by online advertising can be viewed as the outcome of a cooperative effect of the channels and media platforms. Dalessandro et al. [14] pose attribution as a causal estimation problem and propose SV as an approximation scheme for the causally motivated problem. They also show that SV generalizes the probabilistic model of Shao and Li [42]. Using a stylized model, Berman [8] shows the use of SV for attribution can be beneficial to the advertiser.

Other works. Attribution has been tackled from various other angles. Jordan et al. [26] use a Markovian model to motivate a payment scheme that satisfies incentive compatibility for the advertiser and is fair from the publisher's point-of-view. Xu et al. [46] propose a mutually exciting point process to capture dynamic interactions among various ads. Zhang et al. [47], Ji et al. [25], and Ji and Wang [24] provide an interesting view of attribution via the lens of survival theory. Under a stylized setting, Abhishek et al. [1] analyze attribution contracts used by an advertiser to incentivize two publishers that affect customer acquisition. Zhao et al. [48] propose a regression-based relative importance method to compute the marginal contributions.

1.2 Our approach and contributions

With almost a decade of research, it remains unclear as to what is an "appropriate" or "best" attribution measure. IVH appears to be the most popular; but, to the best of our knowledge, there exists little rationale supporting it. On the other hand, SV has a strong theoretical justification. It has a number of desirable properties such as *efficiency*, *symmetry*, *linearity*, and *null player*; in fact, SV is the unique solution to a cooperative game that has all these properties. However, estimating SV exactly is computationally intractable, and one has to resort to approximations [7, 10, 17, 31–34, 36]. In addition, as we show in Section 4.2, SV is not counterfactual in nature. We seek a metric that has the desirable properties of SV, and yet is tractable and able to accommodate counterfactual reasoning.

Our main contributions are as follows. First, we construct an abstract Markov chain model for the user journey through the conversion funnel that generalizes most of the existing Markovian models in the attribution literature. At every period, the user is in one of the finitely many *states*. The advertiser observes the state and takes an *action*. The user transitions to a random state distributed according to a probability mass function that depends on the current state and the advertising action. We propose attributing value to each state-action pair, which is a generalization of the existing approaches that attribute only to advertising actions. This extension allows us to capture state-specific attribution for each action, the need for which is motivated by Bleier and Eisenbeiss [9].

Second, using our Markovian model, we develop a series of intuitive canonical network structures that serve as a robustness check for any attribution scheme. Using these networks, we show the current attribution metrics (LTA, IVH, and uniform) have serious limitations. Furthermore, we show how one can compute state-specific SV for each action in our Markovian model and highlight that it does not adjust for the counterfactual and hence, is an inappropriate metric for attribution. To the best of our knowledge, this is the first work in the literature to analyze the various existing attribution metrics using a common framework.

Third, we develop an axiomatic framework for attribution in online advertising, which forms the main contribution of this work.

[1] Conversion refers to the event in which a user buys the underlying product.

In particular, we propose a new metric for attribution in the Markovian model of user behavior, that we refer to as *counterfactual adjusted Shapley value*. We show that our proposed metric inherits the desirable axioms of the classical SV. We also demonstrate its robustness to a mix of network structures that highlighted limitations of the existing metrics. In addition, we establish that our metric admits a crisp characterization under our Markovian model. It coincides with a *unique-uniform attribution scheme* that explicitly adjusts for the counterfactual. In turn, it can be interpreted as a correction to the commonly used uniform attribution scheme. Furthermore, we exploit this characterization to shed light on the ease of computability of our metric.

The remainder of this paper is organized as follows. In Section 2, we introduce the Markovian model that describes the user journey as a function of ad exposure. We discuss how the existing attribution schemes (LTA, IVH, and uniform) can be captured on the Markovian model and highlight their limitations in Section 3. In Section 4, we showcase the drawback of directly applying SV to measure attribution and then present our novel metric along with the axioms it satisfies. We then characterize our metric as an adjusted unique-uniform attribution scheme in Section 5 and discuss its ease of computability. We conclude in Section 6 where we also discuss some directions for ongoing and future research. Some technical results are deferred to Appendix A.

2 MODEL

We propose a Markovian model for user behavior. The transitions in user's state are stochastic, and are a function of only the current state and the advertiser action. This process ends when the user quits (leaves the system) or converts (buys the product). In Section 2.1, we define the components of the Markov chain (denoted by $\mathcal{M}$). When constructing the Markov chain, we keep most of its elements abstract to showcase its flexibility. In Section 2.2, we shed light on the abstractness through practical examples. We conclude this section by defining the attribution problem over our Markovian model (Section 2.3).

2.1 Markovian model of user behavior

We first discuss the state space of the Markov chain followed by the arrival process of the users. We then elaborate on the action space of the advertiser. Finally, we define the transition probabilities.

State space. We define $\mathbb{S} := \{s\}_{s=1}^{m}$ as the set of states excluding the two absorbing states (quit q and conversion c) and $\mathbb{S}^+ := \mathbb{S} \cup \{q, c\}$. In order to highlight the flexibility of our model, we do not yet give a concrete meaning to a state.

Arrival process. External traffic arrives at state $s \in \mathbb{S}$ w.p. λ_s (*initial state probability*). We define the vector $\lambda \in \mathbb{R}^m$ as $[\lambda_s]_{s\in\mathbb{S}}$. We assume no external traffic arrives at c and q.

Action space. We define $\mathbb{A} := \{a\}_{a=1}^{n}$ as the set of actions the advertiser can take, such as sending an email or showing a display ad. We include the no-ad action ($a = 1$)[2] in $\mathbb{A}$ and define β_s^a to be the probability that an advertiser takes action $a \in \mathbb{A}$ at state $s \in \mathbb{S}$. We denote by β the collection of all β_s^a values and assume it is known.

Transition probabilities. We denote by $p_{ss'}^a$ the probability a user moves from $s \in \mathbb{S}$ to $s' \in \mathbb{S}$ in one transition given the advertiser takes action $a \in \mathbb{A}$ at s. Also, for all $(s, s') \in \mathbb{S}^2$, we define $p_{ss'}^\beta := \sum_{a\in\mathbb{A}} \beta_s^a p_{ss'}^a$, which denotes the *average* transition probability. To keep the notation concise, for each $a \in \mathbb{A}$, we define the matrix $P^a := [p_{ss'}^a]_{(s,s')\in\mathbb{S}^2} \in \mathbb{R}^{m\times m}$. Furthermore, $P^\beta := [p_{ss'}^\beta]_{(s,s')\in\mathbb{S}^2} \in \mathbb{R}^{m\times m}$ and $B^a := \text{diag}([\beta_s^a]_{s\in\mathbb{S}}) \in \mathbb{R}^{m\times m}$ for each $a \in \mathbb{A}$ is a diagonal matrix. Clearly, P^β represents the transition matrix over the *partial* state space $\mathbb{S}$ and $P^\beta = \sum_{a\in\mathbb{A}} B^a P^a$. For all $s \in \mathbb{S}$, we define the vector $p_s^a \in \mathbb{R}^m$ as the s-th row of P^a for all $a \in \mathbb{A}$ and $p_s^\beta \in \mathbb{R}^m$ as the s-th row of P^β. Next, we state the only assumption we make on our problem primitives.

Assumption 2.1. *The Markov chain with P^β as its transition matrix is absorbing, i.e., the probability each user will eventually either quit or convert from any state equals 1.*

We note that Assumption 2.1 is very mild. So far, we have not discussed the transitions to and from states q and c. We use the notation p_{sc}^a and p_{sc}^β to denote the action-specific and average one-step transition probabilities from $s \in \mathbb{S}$ to c for all $a \in \mathbb{A}$. (For transitions from $s \in \mathbb{S}$ to q, replace the index c by q.) Since both q and c are absorbing states, the transitions from them are self-loops w.p. 1. We define $p_c^a := [p_{sc}^a]_{s\in\mathbb{S}} \in \mathbb{R}^m$ for each $a \in \mathbb{A}$ and $p_c^\beta := [p_{sc}^\beta]_{s\in\mathbb{S}} \in \mathbb{R}^m$. (Note that p_s^a and p_s^β for $s \in \mathbb{S}$ correspond to the probabilities of *leaving* s whereas p_c^a and p_c^β correspond to the probabilities of *entering* c.) Next, we shift our focus to three quantities of interest for this Markov chain.

Expected number of visits. We define $F^\beta \in \mathbb{R}^{m\times m}$ as the *fundamental matrix* of $\mathcal{M}$, i.e., its (i, j)-th entry equals the expected number of visits to state j if the initial state is i. Results in Markov chain theory [18] imply that $F^\beta = (I - P^\beta)^{-1}$ and Assumption 2.1 ensures that F^β exists and is finite.

Effective arrival rate. Let μ_s^β denote the *effective* arrival rate into state $s \in \mathbb{S}$ and $\mu^\beta := [\mu_s^\beta]_{s\in\mathbb{S}} \in \mathbb{R}^m$. It is easy to show that $\lambda^\top + (\mu^\beta)^\top P^\beta = (\mu^\beta)^\top$; hence, $(\mu^\beta)^\top = \lambda^\top F^\beta$.

Eventual conversion probability. We define h_s^β as the probability of eventually being absorbed in c from state $s \in \mathbb{S}$ and the vector $h^\beta := [h_s^\beta]_{s\in\mathbb{S}} \in \mathbb{R}^m$. It is easy to show that $h^\beta = P^\beta h^\beta + p_c^\beta$. Thus, $h^\beta = F^\beta p_c^\beta$. We set $h_q^\beta = 0$ and $h_c^\beta = 1$.

2.2 Model discussion

We have so far defined the state and action spaces in abstract terms. Now, we provide some possible mappings from physical quantities to these.

As a first possibility, the state may be composed of past interactions with the user, e.g., the number of visits to the product website, number of emails received, number of emails opened, number of display ads seen, number of display ads clicked, etc. The number of states may become too large, resulting in potential tractability issues. However, one can quantize the counts to control the size of the state space. The advertiser's action space includes doing nothing (no ad), or sending an email, or showing a display ad, etc.

[2] We are reserving $a = 0$ for future use as it will become transparent in Section 4.2.

As another possibility, the state space may correspond to the widespread conversion funnel used in the marketing literature which captures the journey of the user to the state of being unaware of the product to eventually becoming interested and finally purchasing. Given the fine granularity of data available to advertisers, we believe the advertiser can infer such states using an appropriate statistical model. The action space remains the same as above.

We note that the definitions of state and action spaces can be customized as per the needs of the advertiser (since our attribution framework is developed for the abstract Markovian model). In fact, the state space should very much be context-specific and driven by the user features that are relevant to a particular advertiser.

2.3 The attribution problem

The total value[3] generated by the network equals $\lambda^\top h^\beta$. The goal of an attribution model is to allocate this value to the underlying *agents* (state-action pairs)[4] in the system while accounting for the counterfactual. For any given $(s, a) \in \mathbb{S} \times \mathbb{A}$, we use π_s^a to denote the corresponding attribution. Naturally, a desirable property is that the total value in the system should be credited back to the agents, i.e., $\sum_{s\in\mathbb{S}} \sum_{a\in\mathbb{A}} \pi_s^a = \lambda^\top h^\beta$.

3 CURRENT APPROACHES & LIMITATIONS

We now define in our context the commonly used attribution schemes (LTA, IVH, and uniform) and illustrate their limitations. At a high-level, both LTA and uniform are "backward looking" in the sense that they split the value of a path after observing its realization (from the initial state to the end). LTA allocates all the value to the last interacting agent whereas uniform allocates it equally to all the agents that appeared in the path. On the other hand, IVH is "forward looking" since it attributes to a given agent based on what will happen in the future. It does not require the knowledge of how the actual path will unfold. All it requires is a model that can output the change in eventual conversion probability that results if the agent is removed from the path.

3.1 LTA

In LTA, all the value generated due to a purchase is attributed to the last state-action pair in the path. Denote by $\mathcal{P}$ a random path (over state-action pairs) sampled uniformly from $\mathcal{M}$. For $(s, a) \in \mathbb{S} \times \mathbb{A}$, define

$$\hat{w}_s^a(\mathcal{P}) := \begin{cases} 1 & \text{if } \mathcal{P} \text{ converts and } (s,a) \text{ is the last agent in } \mathcal{P} \\ 0 & \text{otherwise.} \end{cases}$$

The attribution to $(s, a) \in \mathbb{S} \times \mathbb{A}$ under the LTA scheme equals

$$\pi_s^{a,\text{LTA}} := \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\hat{w}_s^a(\mathcal{P})\right].$$

In other words, (s, a) receives complete credit of each *converted* path it is the last agent of. Trivially, LTA attributes exactly all the value generated by the network (*efficient*) and is easy to implement. However, LTA appears "unfair" since the states are not rewarded for moving the user up in the conversion funnel. We demonstrate this limitation next.

[3] We assume that the sale of one unit of the product generates a value of 1. If the true number is different from 1, the numbers can be scaled accordingly.

[4] One might want to consider actions as the agents. However, we view the state-action pairs as the agents, because the impact of an action is likely to be state dependent.

Example 3.1 (LTA is unfair). Consider the network in Figure 1 with self-loop probability $p = 0$ (*line*). Under LTA, all the value goes to the ad action at state m although state 1 brings in all the traffic and the ad action at each state pushes the user towards conversion.

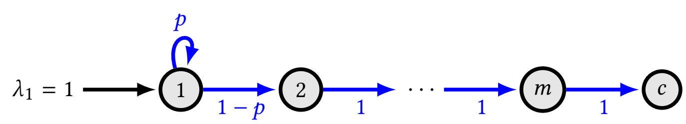

Figure 1: Network for Examples 3.1 and 3.4. The action space consists of two actions: show no-ad and show an ad. The lines denote transitions if an ad is shown. For brevity, we do not show the transitions if an ad is not shown (to quit state w.p. 1). The advertiser shows an ad at all states w.p. 1.

3.2 IVH

IVH allocates to each state-action pair (s, a) the increase in the eventual conversion probability by taking action a in state s as opposed to the no-ad action. For each $a \in \mathbb{A}$, we first define an auxiliary variable that captures the corresponding forward looking increment *conditioned* on the user being at a given state:

$$z^{a,\text{IVH}} := \underbrace{P^a h^\beta + p_c^a}_{\text{action } a} - \underbrace{(P^1 h^\beta + p_c^1)}_{\text{no-ad action}} = \underbrace{(P^a - P^1)h^\beta}_{\text{eventual}} + \underbrace{(p_c^a - p_c^1)}_{\text{immediate}},$$

where $z^{a,\text{IVH}} = [z_s^{a,\text{IVH}}]_{s\in\mathbb{S}} \in \mathbb{R}^m$. The scalar $z_s^{a,\text{IVH}}$ can be interpreted as the allocation at a "trace" level, i.e., if the advertiser observes a user at state s and decides to take action a, the corresponding allocation would be $z_s^{a,\text{IVH}}$. However, we need to scale this metric for it to be seen as attribution over the entire population. In particular, given that at state s, the effective arrival rate is μ_s^β and action a is taken w.p. β_s^a, the IVH attribution is given by

$$\pi_s^{a,\text{IVH}} := \mu_s^\beta \beta_s^a z_s^{a,\text{IVH}}.$$

Although IVH is tractable and somewhat adjusts for the counterfactual, a serious limitation is that it can distribute more value than the network generates because it pays for both eventual and immediate conversions; consequently, each conversion is accounted for several times. We now show that this is, indeed, possible.

Example 3.2 (IVH can over-allocate). Consider the network in Figure 1 with $p = 0$. Under IVH, all states receive an attribution of 1 for the show ad action and 0 for the no-ad action, resulting in a total allocation of m even though the value generated equals 1.

A common workaround to ensure IVH in budget-balanced is to normalize the output by an appropriate constant (*normalized IVH*). Normalizing the numbers in Example 3.2 results in the ad action at each state receiving $1/m$, which appears reasonable. However, as we illustrate in Example 3.3, even normalized IVH can be problematic.

Example 3.3 (Normalized IVH can be unjustified). Consider the network in Figure 2. IVH attributes 1 to state 1 and 1/2 for all other states. Thus, normalized IVH attributes

$$\begin{cases} \frac{1}{1+(m-1)/2} & \text{if state equals 1} \\ \frac{1/2}{1+(m-1)/2} & \text{otherwise.} \end{cases}$$

Clearly, as $m \to \infty$, all the value goes to states 2 to m. This appears inappropriate as state 1 "deserves" at least half the total value since 50% of the users convert immediately after state 1.

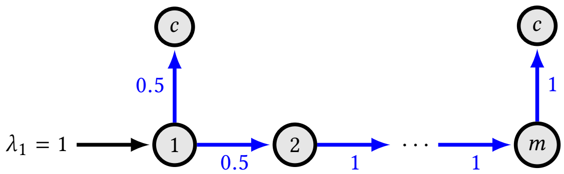

Figure 2: Network for Example 3.3. The action space consists of two actions: show no-ad and show an ad. Solid blue lines denote transitions if an ad is shown. For brevity, we do not show the transitions if an ad is not shown (to quit state w.p. 1). The advertiser shows an ad at all states w.p. 1.

3.3 Uniform attribution

Under a uniform attribution scheme, the credit generated is attributed equally to all the state-action pairs that are encountered in a path. Denote by $\mathcal{P}$ a random path (over state-action pairs) sampled uniformly from $\mathcal{M}$. For $(s, a) \in \mathbb{S} \times \mathbb{A}$, define

$$\bar{w}_s^a(\mathcal{P}) := \begin{cases} \frac{n_s^a}{|\mathcal{P}|} & \text{if } \mathcal{P} \text{ converts and } (s,a) \in \mathcal{P} \\ 0 & \text{otherwise,} \end{cases}$$

where n_s^a equals the number of times (s, a) appears in $\mathcal{P}$ and $|\mathcal{P}|$ denotes the number of state-action pairs in $\mathcal{P}$ (not necessarily unique). The uniform attribution for $(s, a) \in \mathbb{S} \times \mathbb{A}$ is

$$\pi_s^{a,\text{uni}} := \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\bar{w}_s^a(\mathcal{P})\right]. \tag{1}$$

In other words, (s, a) receives an "equal cut" of each *converted* path it contributes to. Such a scheme is efficient and scalable. On the line network, it attributes $1/m$ to the ad action at each state, which seems reasonable. However, it does not account for the counterfactual. Furthermore, it creates undesired incentives, which we show next.

Example 3.4 (Undesired incentive). Consider the network in Figure 1 with $p < 1$. For a given path with n_1 occurrences of state 1, the uniform attribution scheme attributes $n_1/(n_1 + m - 1)$ to state 1. As p increases, state 1 receives more credit. However, this seems inappropriate since it gives an undesired incentive to the ad action at state 1 to "game the system" via self-loops. Ideally, an attribution scheme should incentivize the ads to push the user towards conversion.

Remark 3.1 (Unique-uniform). A simple fix to the above drawback is to attribute based on the number of unique agents. For instance, in the example above, the *unique-uniform* scheme would attribute $1/m$ to the ad action at each state. However, this scheme still does not account for the counterfactual, and there is no formal rationale for it. (In the following sections, we will see that adjusting for uniqueness is in fact supported by a strong mathematical rationale and will be part of our prescribed method.)

To summarize this section, the above examples highlight some limitations of existing heuristics but also point to the fact that no single heuristic "dominates the other." In some way, IVH possesses a counterfactual form giving it some practical appeal. At the same time, it is only forward looking, ignoring all past actions and their potential contributions. Uniform, on the other hand, accounts for past actions, but is not counterfactual. Ideally, an attribution scheme would provide the best of both worlds, properly accounting for past actions while also accounting for the counterfactual associated with not advertising.

4 SHAPLEY VALUE

We now study SV as a measure for attribution. We first present a primer on SV (Section 4.1) followed by a discussion on why a direct application to our context does not suffice (Section 4.2). We then adapt SV accordingly (Section 4.3) and conclude this section by discussing the performance of our proposed measure on all the motivating examples (Section 4.4). In this section, we hope to convince the reader that, leaving computational tractability aside, SV (adjusted for the counterfactual) is an appropriate metric for attribution.

4.1 Primer

SV is a well-accepted concept from cooperative game theory that is used to distribute the payoff generated by a *coalition* to the *players* in the coalition [43]. Given a finite set $\mathbb{P}$ of players, the *characteristic function* $v(\cdot)$ maps a coalition $\mathcal{X} \subseteq \mathbb{P}$ to the value (a real number) generated by the coalition. The value $v(\emptyset)$ of the empty coalition is normalized to 0. SV distributes the value $v(\mathbb{P})$ of the grand coalition to a player $r \in \mathbb{P}$ as follows:

$$\pi_r^{\text{Shap}} := \sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|} \times \{v(\mathcal{X}\cup\{r\}) - v(\mathcal{X})\},$$

where

$$w_{|\mathcal{X}|} := \frac{|\mathcal{X}|!(|\mathbb{P}| - |\mathcal{X}| - 1)!}{|\mathbb{P}|!}.$$

The attractiveness of SV is rooted in the fact that it is the *only* solution to a cooperative game that has the following four desirable properties:

1. **Efficiency**: $\sum_{r\in\mathbb{P}} \pi_r = v(\mathbb{P})$.
2. **Symmetry**: Consider players $r, r' \in \mathbb{P}$ such that for any $\mathcal{X} \subseteq \mathbb{P} \setminus \{r, r'\}$, r and r' are *equivalent*, i.e., $v(\mathcal{X} \cup \{r\}) = v(\mathcal{X} \cup \{r'\})$. Then, $\pi_r = \pi_{r'}$.
3. **Linearity**: Consider two characteristic functions $v(\cdot)$ and $w(\cdot)$. Linearity states that for all players $r \in \mathbb{P}$, $\pi_r(v + w) = \pi_r(v) + \pi_r(w)$ and $\pi_r(\alpha v) = \alpha\pi_r(v)$ for all $\alpha \in \mathbb{R}$.
4. **Null player**: Suppose player $r \in \mathbb{P}$ does not add any value to any coalition, i.e., for all $X \subseteq \mathbb{P}$, $v(X \cup \{r\}) = v(X)$. Then, $\pi_r = 0$.

4.2 Direct application

SV is a natural candidate for decomposing network value because the state-action pairs can be viewed as players participating in a cooperative game to achieve a common goal of converting the users. In particular, given the action intensities β, the network generates a value of $\lambda^\top h^\beta$ and the state-action specific SV $\pi_s^{a,\text{Shap}}$ represents the credit allocated to each state-action pair $(s, a) \in \mathbb{S}\times\mathbb{A}$. We first define the underlying components (players, coalition, and characteristic function) in our context and then discuss an important drawback of such a naive application.

Player. We view each state-action pair $(s, a) \in \mathbb{S} \times \mathbb{A}$ as an underlying player in the cooperative game. Consequently, our SV will not just be action-specific, but also state-specific.

Coalition. A coalition $\mathcal{X} \subseteq \mathbb{S} \times \mathbb{A}$ refers to a collection of players (state-action pairs). We construct a coalition by starting with the *empty coalition*, which corresponds to the Markov chain containing all the states but no action (not even the no-ad action), i.e., the traffic is directed to the quit state w.p. 1. Equivalently, one can view the empty coalition as the one with all the states but *action 0* ("zero-value action") being taken at all of them[5]. If a state-action pair, say (s, a), is added to the empty coalition, then the advertiser takes action a at state s w.p. β_s^a and so on. Therefore, for each coalition $\mathcal{X}$, we obtain a corresponding β, which we denote by $\beta^{\mathcal{X}}$.

Characteristic function. Given $\mathcal{X} \subseteq \mathbb{S} \times \mathbb{A}$, we define

$$v(\mathcal{X}) := \lambda^\top h^{\beta^{\mathcal{X}}}.$$

We define the action probabilities collections $\beta^{\emptyset}$ and β^* such that they correspond to the empty coalition $\emptyset$ and complete coalition $\mathbb{S} \times \mathbb{A}$, respectively. Thus, $\beta^{\emptyset} = 0$ and $\beta^* = \{\beta_s^a\}_{(s,a) \in \mathbb{S} \times \mathbb{A}}$. We satisfy $v(\emptyset) = 0$ and $v(\mathbb{S} \times \mathbb{A}) = \lambda^\top h^{\beta^*}$ is the total network value.

Shapley value. The SV for each $(s, a) \in \mathbb{S} \times \mathbb{A}$ is

$$\pi_s^{a,\text{Shap}} := \sum_{\mathcal{X} \subseteq \{\mathbb{S} \times \mathbb{A}\} \setminus \{(s,a)\}} w_{|\mathcal{X}|} \times \{v(\mathcal{X} \cup \{(s, a)\}) - v(\mathcal{X})\}, \quad (2)$$

where

$$w_{|\mathcal{X}|} := \frac{|\mathcal{X}|!(mn - |\mathcal{X}| - 1)!}{(mn)!}. \quad (3)$$

SV depends on our choice of the characteristic function $v(\cdot)$ and the underlying Markov chain $\mathcal{M}$ and hence, we will use the notation $\pi_s^{a,\text{Shap}}(v)$ and $\pi_s^{a,\text{Shap}}(\mathcal{M})$ when the emphasis on $v(\cdot)$ and $\mathcal{M}$ is necessary. Though this view of attribution inherits the desirability of SV, it suffers from a critical flaw: it does not adjust for the counterfactual (Example 4.1).

Example 4.1 (Need for counterfactual). Consider the network in Figure 3. Clearly, showing an ad should not get any attribution since it provides no *additional* value over the counterfactual action (no ad). However, (2) attributes all the value to the ad action. Furthermore, even LTA and uniform attribution fail this sanity check.

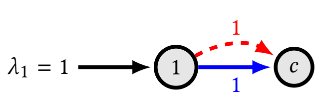

Figure 3: Network for Example 4.1. The action space consists of two actions: show no-ad and show an ad. Solid blue (dashed red) lines denote transition if an ad is shown (not shown). The advertiser shows an ad w.p. 1.

[5]An alternate candidate for the empty coalition is to have the no-ad action being taken at all the states. In other words, the no-ad action fills the "void" left due to the absence of other actions. However, we believe such a view is inappropriate since the no-ad action itself can have some value (either in terms of network effect or in terms of direct conversions). Therefore, using the no-ad action as the "default" action would allow the other actions to free ride on the value generated by the no-ad action, and hence, undermine the contribution due to the no-ad action.

4.3 Our approach

As we alluded to earlier, we seek a measure that provides the best of two worlds: (1) appropriately captures the contributions of the past actions (capturing the feature of, e.g., uniform and SV)[6] and (2) exhibits counterfactual reasoning (capturing a feature of IVH). To this end, we focus on adapting SV to account for the counterfactual and we do so by adhering to the causality framework of Rubin [41]. We first define a *counterfactual player* followed by our novel definition of *counterfactual adjusted Shapley value (CASV)*. We then show the desirability of CASV.

Counterfactual player. Following Rubin [41], the counterfactual to taking action a at state s w.p. β_s^a is to take the no-ad action (action 1) at state s w.p. β_s^a (instead of β_s^1). Accordingly, for a given player $(s, a) \in \mathbb{S} \times \mathbb{A}$, we denote the counterfactual player as $(s, 1)^a$, where the "a" in the superscript captures the dependence on β_s^a.

Counterfactual adjusted Shapley value. The game theoretic setup (player, coalition, characteristic function) is the same as in Section 4.2. For each $(s, a) \in \mathbb{S} \times \mathbb{A}$, we define CASV as

$$\psi_s^{a,\text{Shap}} := \sum_{\mathcal{X} \subseteq \{\mathbb{S} \times \mathbb{A}\} \setminus \{(s,a)\}} w_{|\mathcal{X}|} \times \{v(\mathcal{X} \cup \{(s, a)\}) - v(\mathcal{X} \cup \{(s, 1)^a\})\}, \quad (4)$$

where $w_{|\mathcal{X}|}$ is the same as in (3). We use the symbol ψ instead of π to differentiate CASV from SV. The only change we make in going from $\pi_s^{a,\text{Shap}}$ to $\psi_s^{a,\text{Shap}}$ is that we replace $v(\mathcal{X})$ by $v(\mathcal{X} \cup \{(s, 1)^a\})$, i.e., CASV captures the value added by a player over its counterfactual. Intuitively, it is unclear whether CASV under the current cooperative game is equivalent to SV under a different cooperative game. However, it is possible to view CASV as the difference between two SVs:

$$\psi_s^{a,\text{Shap}}(\mathcal{M}) = \pi_s^{a,\text{Shap}}(\mathcal{M}) - \pi_s^{a,\text{Shap}}(\mathcal{M}_s^a), \quad (5)$$

where $\mathcal{M}$ denotes the original Markov chain and $\mathcal{M}_s^a$ denotes the *counterfactual network* for (s, a), i.e., it is identical to $\mathcal{M}$ except that we replace the transition probabilities of (s, a) by those of $(s, 1)$. Clearly, $\pi_s^{a,\text{Shap}}(\mathcal{M}_s^a)$ is the *counterfactual value* of (s, a), i.e., the value generated if no-ad was shown instead of ad a at state s.

Obviously, it is of interest to check how CASV performs with respect to the desirable properties that motivated SV in the first place. However, one needs to be careful when discussing these axioms in the counterfactual context. To be specific, efficiency should now pertain to the redistribution of additional value generated over the counterfactual value. Similarly, the definition of equivalent players should be adjusted (discussed in Remark 4.1) and null player should correspond to a player with zero value-add. We formalize the desirability of CASV in Theorem 4.2, the proof of which is in Appendix A.

Theorem 4.2 (Axioms). *CASV satisfies the following axioms:*

1. ***Counterfactual efficiency**: The sum of CASVs equals the additional value generated over the counterfactual value, i.e.,*

$$\sum_{(s,a) \in \mathbb{S} \times \mathbb{A}} \psi_s^{a,\text{Shap}} = v(\mathbb{S} \times \mathbb{A}) - \sum_{(s,a) \in \mathbb{S} \times \mathbb{A}} \pi_s^{a,\text{Shap}}(\mathcal{M}_s^a).$$

[6]We will see in Section 5 that uniform and SV are closely linked.

2. ***Counterfactual symmetry***: *If $(s, a) \in \mathbb{S} \times \mathbb{A}$ and $(s', a') \in \mathbb{S} \times \mathbb{A}$ are counterfactual equivalent, i.e.,*
(A) $v(\mathcal{X} \cup \{(s,a)\}) - v(\mathcal{X} \cup \{(s,1)^a\}) = v(\mathcal{X} \cup \{(s',a')\}) - v(\mathcal{X} \cup \{(s',1)^{a'}\})$ *and*
(B) $v(\mathcal{X} \cup \{(s,1)^a, (s',a')\}) = v(\mathcal{X} \cup \{(s',1)^{a'}, (s,a)\})$
for all $\mathcal{X} \subseteq \{\mathbb{S} \times \mathbb{A}\} \setminus \{(s,a),(s',a')\}$, then
$$\psi_s^{a,\text{Shap}} = \psi_{s'}^{a',\text{Shap}}.$$
3. ***Linearity***: *Consider two characteristic functions $v(\cdot)$ and $w(\cdot)$. For all $(s,a) \in \mathbb{S} \times \mathbb{A}$, we have*
$$\psi_s^{a,\text{Shap}}(v+w) = \psi_s^{a,\text{Shap}}(v) + \psi_s^{a,\text{Shap}}(w)$$
and for all $\alpha \in \mathbb{R}$,
$$\psi_s^{a,\text{Shap}}(\alpha v) = \alpha \psi_s^{a,\text{Shap}}(v).$$
4. ***Counterfactual null player***: *Consider a player $(s,a) \in \mathbb{S} \times \mathbb{A}$ that has a zero value-add to all coalitions that do not contain (s,a), i.e., for all $\mathcal{X} \subseteq \{\mathbb{S} \times \mathbb{A}\} \setminus \{(s,a)\}$,*
$$v(\mathcal{X} \cup \{(s,a)\}) = v(\mathcal{X} \cup \{(s,1)^a\}).$$
Then, $\psi_s^{a,\text{Shap}} = 0$.

Furthermore, CASV is the unique solution to satisfy these four counterfactual axioms.

Remark 4.1 (Counterfactual equivalent players). In Theorem 4.2, we define $(s,a) \in \mathbb{S} \times \mathbb{A}$ and $(s',a') \in \mathbb{S} \times \mathbb{A}$ to be counterfactual equivalent if they satisfy (A) and (B). To develop an intuitive understanding of these conditions, one can consider the following three conditions for all $\mathcal{X} \subseteq \{\mathbb{S} \times \mathbb{A}\} \setminus \{(s,a),(s',a')\}$:

(A1) $v(\mathcal{X} \cup \{(s,a)\}) = v(\mathcal{X} \cup \{(s',a')\})$,
(A2) $v(\mathcal{X} \cup \{(s,1)^a\}) = v(\mathcal{X} \cup \{(s',1)^{a'}\})$, and
(B) $v(\mathcal{X} \cup \{(s,1)^a, (s',a')\}) = v(\mathcal{X} \cup \{(s',1)^{a'}, (s,a)\})$.

In simple words, consider a coalition that does not contain either of the two given players. Condition (A1) states that adding either of the two players has the same effect on the network value and condition (A2) says so for the counterfactual players. The remaining case of interest is what happens when, on one hand, we add the first player and the counterfactual of the second player whereas, on the other hand, we do the opposite (add the second player and the counterfactual of the first player). Condition (B) states that doing either is equivalent in terms of the characteristic function. Note that we only need conditions (A) and (B) for two players to be counterfactual equivalent, and (A1) and (A2) are sufficient conditions for (A). It is worth mentioning that our definition of counterfactual equivalent players is a generalization of equivalent players from Section 4.1. In particular, equivalent players only need to satisfy condition (A1) and since the notion of a counterfactual player does not exist in the SV context, conditions (A2) and (B) are irrelevant.

We conclude this subsection with a discussion regarding the attribution to the no-ad action. By definition of CASV, the no-ad action at each state receives zero counterfactual credit. However, one can attribute the counterfactual value $\sum_{(s,a)\in\mathbb{S}\times\mathbb{A}} \pi_s^{a,\text{Shap}}(\mathcal{M}_s^a)$ ("residual") to the no-ad action in a post hoc fashion. In other words, the post hoc attribution to the no-ad action at state $s \in \mathbb{S}$ equals
$$\sum_{a\in\mathbb{A}} \pi_s^{a,\text{Shap}}(\mathcal{M}_s^a) = \pi_s^{1,\text{Shap}}(\mathcal{M}) + \sum_{a=2}^{n} \pi_s^{a,\text{Shap}}(\mathcal{M}_s^a), \tag{6}$$
where we use the fact that $\mathcal{M}_s^1 = \mathcal{M}$. The two terms in (6) have an intuitive interpretation. The first term captures the value generated by the no-ad action due to a positive value of β_s^1 whereas the second term captures the value that *would have* been generated if the no-ad action was used in place of the other actions.

4.4 Revisiting canonical examples

We now revisit the networks discussed so far and show that our CASV measure does not suffer from the same limitations as existing heuristics.

Example 4.3. For the network in Figure 1 with $p < 1$, CASV at every state equals $1/m$ for the ad action and 0 otherwise.

Example 4.4. For the network in Figure 2, CASV attributes 0 to the no-ad action at all the states and the following to the ad action:
$$\begin{cases} \frac{1}{2} + \frac{1}{2m} & \text{if } s = 1 \\ \frac{1}{2m} & \text{otherwise.} \end{cases}$$
This seems sensible as half of the paths convert just due to the ad action at state 1 whereas the other half of the paths convert with the help of the ad action at all the m states.

Example 4.5. For the network in Figure 3, CASV allocates zero credit to the ad action, which seems appropriate. However, CASV of the no-ad action also equals zero, which raises the following question: which player gets credit for the value in the system? The answer is, as discussed at the end of Section 4.3, the no-ad action.

In sum, CASV appears to be an appealing measure for attribution: it has a number of desirable properties and appears robust to various network structures. However, similar to SV, computing CASV using (4) requires an exponential runtime in the number of underlying players. This is an important tractability concern that could render CASV impractical in settings with moderately-sized state and action spaces. In the next section, we focus on representing CASV in a manner that is amenable to being computed efficiently, under our Markovian model.

5 CHARACTERIZING COUNTERFACTUAL ADJUSTED SHAPLEY VALUE

We now focus on characterizing CASV by exploiting the structure of the underlying cooperative game. In doing so, we use the fact that CASV can be expressed as a difference of two SVs (see (5)) and hence, analyze SV first. Proposition 5.1 connects the coalition-oriented game-theoretic construct of SV to the paths sampled from $\mathcal{M}$. In particular, we show that, under our Markovian setup, SV is (surprisingly) identical to the unique-uniform attribution scheme (motivated in Remark 3.1). We present the formal proof of Proposition 5.1 in Appendix A and highlight the intuition here.

Proposition 5.1 (SV equals unique-uniform). *Consider $(s,a) \in \mathbb{S} \times \mathbb{A}$ and the Markov chain $\mathcal{M}$. The SV of (s,a) as defined in* (2) *equals*

$$\pi_s^{a,\,Shap} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[w_s^a(\mathcal{P})\right],$$

where

$$w_s^a(\mathcal{P}) := \begin{cases} \frac{1}{u(\mathcal{P})} & \text{if } \mathcal{P} \text{ converts and } (s,a) \in \mathcal{P} \\ 0 & \text{otherwise.} \end{cases}$$

The function $u(\cdot)$ returns the number of unique players and $\mathcal{P}$ is a path over players, i.e., state-action pairs.

Proof sketch. There are two keys ideas at interplay here. First, to characterize SV, we observe that the characteristic function for a given coalition $\mathcal{X} \subseteq \mathbb{S}\times\mathbb{A}$ can be expressed as an expectation over the paths sampled from the Markov chain $\mathcal{M}(\mathcal{X})$ in which only the players in $\mathcal{X}$ are "active", i.e.,

$$v(\mathcal{X}) = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}(\mathcal{X})}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}}\right].$$

Using this observation allows us to express SV $\pi_s^{a,\,\text{Shap}}$ as an expectation over the paths from the original Markov chain $\mathcal{M}$:

$$\pi_s^{a,\,\text{Shap}} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}}\,(\star)\right],$$

where

$$(\star) := \sum_{\mathcal{X}\subseteq\{\mathbb{S}\times\mathbb{A}\}\setminus\{(s,a)\}} w_{|\mathcal{X}|}\left(\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{(s,a)\}\}} - \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right).$$

The second key idea is to analyze $(\star)$ rigorously, which forms the bulk of our analysis. We show that $(\star)$ admits the following remarkably simple form:

$$(\star) = \begin{cases} \frac{1}{u(\mathcal{P})} & \text{if } (s,a)\in\mathcal{P} \\ 0 & \text{otherwise.} \end{cases}$$

Putting the two pieces together allows us to finish the proof. □

The above result enables one to characterize CASV, which constitutes the main result of this work.

Theorem 5.1 (CASV equals unique-uniform minus counterfactual). *Consider $(s,a)\in\mathbb{S}\times\mathbb{A}$, the Markov chain $\mathcal{M}$, and the counterfactual Markov chain $\mathcal{M}_s^a$. The CASV of (s,a) as defined in* (4) *equals*

$$\psi_s^{a,\,Shap} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[w_s^a(\mathcal{P})\right] - \mathbb{E}_{\mathcal{P}\sim\mathcal{M}_s^a}\left[w_s^a(\mathcal{P})\right],$$

where $w_s^a(\cdot)$ and $u(\cdot)$ are as defined in Proposition 5.1.

There is a striking resemblance between this characterization of CASV and uniform attribution as defined in (1). In particular, the two expressions are identical except the definition of the weight function and the counterfactual adjustment. The weight function in CASV does not provide an incentive for players to "game the system" as it only rewards based on whether a player appears in the path or not (as opposed to the number of times it appears). In other words, if an ad a had to be shown multiple times (say n_s^a) at the *same* state s to make the user move to another state s', then the CASV weight function only rewards it once (as opposed to rewarding it n_s^a times). This seems very reasonable and, in fact, is the simple fix we proposed in Remark 3.1.

It is noteworthy that the coalition-oriented construct of CASV, which on the surface does not seem to be related to the paths of the underlying Markovian model, reduces to being expressed as a remarkably simple function of such paths. This connection is quite valuable as it helps to gain deeper insights regarding the structure of CASV and hence, build a better understanding. Next, we conclude this section by using this connection to highlight that our scheme is efficiently computable and implementable.

Given a set of user paths $\mathcal{D} := \{\mathcal{P}_i\}_{i=1}^N$ where each user path $\mathcal{P}_i$ consists of various state-action-state (s,a,s') tuples, the action-specific transition probabilities can be estimated using empirical frequencies. To estimate CASV for player (s,a), one can sample paths from the estimated Markov chains $\mathcal{M}$ and $\mathcal{M}_s^a$ and use Theorem 5.1 directly. However, this involves sampling from a different Markov chain $\mathcal{M}_s^a$ to estimate CASV for each player, which we address next.

A simple change of measure allows CASV to be expressed as

$$\psi_s^{a,\,\text{Shap}} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[w_s^a(\mathcal{P})\left(1 - \frac{g_s^a(\mathcal{P})}{g(\mathcal{P})}\right)\right], \tag{7}$$

where $g(\mathcal{P})$ and $g_s^a(\mathcal{P})$ denote the probabilities of observing path $\mathcal{P}$ under Markov chains $\mathcal{M}$ and $\mathcal{M}_s^a$, respectively. The ratio $g_s^a(\mathcal{P})/g(\mathcal{P})$ denotes the *importance weight* and it is easy to show that

$$\frac{g_s^a(\mathcal{P})}{g(\mathcal{P})} = \prod_{s'\in\mathbb{S}^+}\left(\frac{p_{ss'}^1}{p_{ss'}^a}\right)^{n_{ss'}^a(\mathcal{P})},$$

where $n_{ss'}^a(\mathcal{P})$ denotes the number of occurrences of the tuple (s,a,s') in $\mathcal{P}$ for each $s'\in\mathbb{S}^+$. Given (7), it suffices to sample paths from just one Markov chain ($\mathcal{M}$) to estimate CASV for all the players. In fact, each sample can be recycled for multiple players and the scheme can be implemented using a parallel (over both players and samples) architecture.

It is possible to use a Bayesian version of the approach described above. In particular, one can maintain a belief over the transition probability vector $[p_{ss'}^a]_{s'\in\mathbb{S}^+}$ in the form of a Dirichlet distribution for each $a\in\mathbb{A}$. Using Dirichlet-multinomial conjugacy and real data $\mathcal{D}$, the belief can be updated easily. The posterior belief can be used to generate samples of the Markov chain $\mathcal{M}$ itself. For each sample of $\mathcal{M}$, one can use (7) to estimate CASV, resulting in posterior samples of CASV. Naturally, these posterior samples quantify the uncertainty in CASV that arises due to the noise in data and/or lack of data. Furthermore, this Bayesian approach is amenable to parallelization over the samples of $\mathcal{M}$.

6 CONCLUSIONS & FURTHER RESEARCH

In this paper, using a Markovian model for user behavior, we have proposed a new metric (counterfactual adjusted Shapley value) for the attribution problem in online advertising. We have established its grounding and appropriateness in two ways. First, we have shown the robustness of the proposed measure over various canonical settings in which the existing metrics fail. Second, we have provided an underlying axiomatic framework motivated by game theory and causality that supports our choice. Furthermore, we have established a characterization of the proposed metric as a remarkably simple function of the user paths. In particular, we have established, in our Markovian model, that the proposed attribution metric is an adjustment to the unique-uniform attribution scheme. Finally, we have proposed multiple approaches to efficiently estimate our metric.

There are many potential directions for future research, most of which we are pursuing at the present. First, it is of interest to develop more canonical settings to further verify the appropriateness of our proposed measure. Second, though we have briefly touched upon algorithms to estimate the proposed metric, a thorough treatment would involve understanding their statistical efficiency. In addition, it is worthwhile to explore algorithms to estimate the counterfactual adjusted Shapley value that can be implemented in an online fashion while maintaining strong approximation guarantees. Third, with the help of an industry partner, we are working on comparing the output of our scheme to the existing ones on a real world dataset composed of millions of user paths. Fourth, we are interested in applying the methodology developed in this paper to other domains. In particular, the abstract Markovian model presented here can be used to model other applications, where one wishes to understand the contributions of the underlying agents. Finally, it is worthwhile to explore whether attribution metrics can be constructed from other notions of causality. In this work, we modified Shapley value using Rubin's definition of causality [41]. However, there exist several alternative formulations for causality. See, e.g. Chockler and Halpern [11], Collins et al. [13], Eells [16], Halpern and Pearl [20], Hitchcock [21], Morgan and Winship [35], Pearl [37] and Hume [22] for approaches to causality in the computer science and philosophy literature. Chockler and Halpern [11] proposed metrics such as "degree of responsibility" and "blame" in order to *quantify* the causal effect of one variable on another. However, computing these metrics is computationally expensive. Furthermore, these metrics lack axiomatic support. In particular, similar to IVH, these metrics are not budget balanced. In fact, Chockler and Halpern [11] mention adjusting their notions using Shapley value as a potential future work. Accordingly, though we have unified the existing attribution literature in this work, there is more to explore in terms of the set of counterfactuals to consider.

ACKNOWLEDGMENTS

We acknowledge the support of Adobe Digital Marketing Research Award for this project. V. Goyal also acknowledges the support of NSF grants CMMI 1636046 and CMMI 1351838 for this work. We also thank Abhishek Pani, Chen Dong, Manoj Ravi, and Jie Zhang from the Media and Advertising Solutions team at Adobe for useful discussions during the course of this work. Finally, we thank the WWW reviewers for the valuable feedback and pointing us to the responsibility and blame literature.

A TECHNICAL RESULTS

Theorem 4.2 (Axioms). *CASV satisfies the following axioms:*

1. ***Counterfactual efficiency**: The sum of CASVs equals the additional value generated over the counterfactual value, i.e.,*
$$\sum_{(s,a)\in\mathbb{S}\times\mathbb{A}} \psi_s^{a,\,Shap} = v(\mathbb{S}\times\mathbb{A}) - \sum_{(s,a)\in\mathbb{S}\times\mathbb{A}} \pi_s^{a,\,Shap}(\mathcal{M}_s^a).$$
2. ***Counterfactual symmetry**: If $(s,a)\in\mathbb{S}\times\mathbb{A}$ and $(s',a')\in\mathbb{S}\times\mathbb{A}$ are counterfactual equivalent, i.e.,*
   (A) $v(\mathcal{X}\cup\{(s,a)\}) - v(\mathcal{X}\cup\{(s,1)^a\}) = v(\mathcal{X}\cup\{(s',a')\}) - v(\mathcal{X}\cup\{(s',1)^{a'}\})$ *and*
   (B) $v(\mathcal{X}\cup\{(s,1)^a,(s',a')\}) = v(\mathcal{X}\cup\{(s',1)^{a'},(s,a)\})$
   for all $\mathcal{X}\subseteq\{\mathbb{S}\times\mathbb{A}\}\setminus\{(s,a),(s',a')\}$, then
$$\psi_s^{a,\,Shap} = \psi_{s'}^{a',\,Shap}.$$
3. ***Linearity**: Consider two characteristic functions $v(\cdot)$ and $w(\cdot)$. For all $(s,a)\in\mathbb{S}\times\mathbb{A}$, we have*
$$\psi_s^{a,\,Shap}(v+w) = \psi_s^{a,\,Shap}(v) + \psi_s^{a,\,Shap}(w)$$
and for all $\alpha\in\mathbb{R}$,
$$\psi_s^{a,\,Shap}(\alpha v) = \alpha\psi_s^{a,\,Shap}(v).$$
4. ***Counterfactual null player**: Consider a player $(s,a)\in\mathbb{S}\times\mathbb{A}$ that has a zero value-add to all coalitions that do not contain (s,a), i.e., for all $\mathcal{X}\subseteq\{\mathbb{S}\times\mathbb{A}\}\setminus\{(s,a)\}$,*
$$v(\mathcal{X}\cup\{(s,a)\}) = v(\mathcal{X}\cup\{(s,1)^a\}).$$
Then, $\psi_s^{a,\,Shap}=0$.

Furthermore, CASV is the unique solution to satisfy these four counterfactual axioms.

Proof. Counterfactual efficiency and linearity follow from (5) when used with the efficiency and linearity of SV, respectively. Counterfactual null player follows from (4). For counterfactual symmetry, consider $(s,a)\in\mathbb{S}\times\mathbb{A}$ and $(s',a')\in\mathbb{S}\times\mathbb{A}$ satisfying (A) and (B), define the set $\diamond := \{\mathbb{S}\times\mathbb{A}\}\setminus\{(s,a),(s',a')\}$, and observe that

$$\begin{aligned}
\psi_s^{a,\,\mathrm{Shap}} &= \sum_{\mathcal{X}\subseteq\{\mathbb{S}\times\mathbb{A}\}\setminus\{(s,a)\}} w_{|\mathcal{X}|}\times\{v(\mathcal{X}\cup\{(s,a)\}) - v(\mathcal{X}\cup\{(s,1)^a\})\} \\
&= \sum_{\mathcal{X}\subseteq\diamond} w_{|\mathcal{X}|}\times\{v(\mathcal{X}\cup\{(s,a)\}) - v(\mathcal{X}\cup\{(s,1)^a\})\} \\
&\quad + \sum_{\mathcal{X}\subseteq\diamond} w_{|\mathcal{X}|+1}\times\{v(\mathcal{X}\cup\{(s,a),(s',a')\}) - v(\mathcal{X}\cup\{(s,1)^a,(s',a')\})\} \\
&= \sum_{\mathcal{X}\subseteq\diamond} w_{|\mathcal{X}|}\times\left\{v(\mathcal{X}\cup\{(s',a')\}) - v(\mathcal{X}\cup\{(s',1)^{a'}\})\right\} \\
&\quad + \sum_{\mathcal{X}\subseteq\diamond} w_{|\mathcal{X}|+1}\times\left\{v(\mathcal{X}\cup\{(s,a),(s',a')\}) - v(\mathcal{X}\cup\{(s',1)^{a'},(s,a)\})\right\} \\
&= \sum_{\mathcal{X}\subseteq\{\mathbb{S}\times\mathbb{A}\}\setminus\{(s',a')\}} w_{|\mathcal{X}|}\times\left\{v(\mathcal{X}\cup\{(s',a')\}) - v(\mathcal{X}\cup\{(s',1)^{a'}\})\right\} \\
&= \psi_{s'}^{a',\,\mathrm{Shap}}.
\end{aligned}$$

This proves that CASV satisfies the four axioms. To be concise, we defer the uniqueness proof to the journal version of this paper. □

Proposition 5.1 (SV equals unique-uniform). *Consider $(s,a)\in\mathbb{S}\times\mathbb{A}$ and the Markov chain $\mathcal{M}$. The SV of (s,a) as defined in (2) equals*
$$\pi_s^{a,\,Shap} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[w_s^a(\mathcal{P})\right],$$
where
$$w_s^a(\mathcal{P}) := \begin{cases}\frac{1}{u(\mathcal{P})} & \text{if } \mathcal{P} \text{ converts and } (s,a)\in\mathcal{P}\\ 0 & \text{otherwise.}\end{cases}$$
The function $u(\cdot)$ returns the number of unique players and $\mathcal{P}$ is a path over players, i.e., state-action pairs.

Proof. For convenience, we use linearized notation such that $r := (s, a)$ and $\mathbb{P} := \mathbb{S}\times\mathbb{A}$ with the understanding that $\pi_r^{\text{Shap}} := \pi_s^{a,\text{Shap}}$, and $w_r(\cdot) := w_s^a(\cdot)$. The proof is split into three parts.

Step 1: Express $v(\mathcal{X})$ as an expectation over paths:

$$\begin{aligned}
v(\mathcal{X}) &= \lambda^\top h^{\beta^{\mathcal{X}}} \\
&= \lambda^\top \left(I + P^{\beta^{\mathcal{X}}} + (P^{\beta^{\mathcal{X}}})^2 + \dots\right) p_c^{\beta^{\mathcal{X}}} \\
&= \mathbb{E}_{\mathcal{P}\sim\mathcal{M}(\mathcal{X})}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}}\right] \\
&= \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}}\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right],
\end{aligned}$$

where $\mathcal{M}(\mathcal{X})$ denotes the Markov chain in which only the players in coalition $\mathcal{X}$ are "active".

Step 2: Use the representation from Step 1 in the definition of SV π_r^{Shap} as stated in (2):

$$\begin{aligned}
\pi_r^{\text{Shap}} &= \sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|} \times \{v(\mathcal{X}\cup\{r\}) - v(\mathcal{X})\} \\
&= \sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|} \left\{\mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}}\left(\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}} - \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right)\right]\right\} \\
&= \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[\mathbb{I}_{\{\mathcal{P}\text{ converts}\}} \sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|}\left(\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}} - \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right)\right].
\end{aligned}$$

Step 3: Use Lemma A.1, which states

$$\sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|}\left(\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}} - \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right) = \begin{cases}\frac{1}{u(\mathcal{P})} & \text{if } r\in\mathcal{P} \\ 0 & \text{otherwise,}\end{cases}$$

to conclude that

$$\pi_r^{\text{Shap}} = \mathbb{E}_{\mathcal{P}\sim\mathcal{M}}\left[w_r(\mathcal{P})\right]. \tag{8}$$

This completes the proof. ◻

Lemma A.1 (Unique players). *For any player $r \in \mathbb{P}$ and any path $\mathcal{P}$ over players in $\mathbb{P}$, we have*

$$\sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|}\left(\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}} - \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}\right) = \begin{cases}\frac{1}{u(\mathcal{P})} & \textit{if } r\in\mathcal{P} \\ 0 & \textit{otherwise,}\end{cases}$$

where the function $u(\cdot)$ returns the number of unique players.

Proof. First, suppose $r \notin \mathcal{P}$. Then, $\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}} = \mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}}$ for all $\mathcal{X} \subseteq \mathbb{P}\setminus\{r\}$ and hence, the l.h.s. equals 0, which equals the r.h.s.

Second, suppose $r \in \mathcal{P}$. Assume w.l.o.g. that path $\mathcal{P}$ has $u + 1$ unique players: $r, r_1, \dots, r_u$. Observe that $\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\}} = 0$ for all $\mathcal{X} \subseteq \mathbb{P}\setminus\{r\}$ and hence, the l.h.s. equals $\sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|}\mathbb{I}_{\{\mathcal{P}\subseteq\mathcal{X}\cup\{r\}\}}$. Furthermore, since $r \in \mathcal{P}$, we can define a new path $\mathcal{P}' := \mathcal{P}\setminus\{r\}$ and express the l.h.s. as $\sum_{\mathcal{X}\subseteq\mathbb{P}\setminus\{r\}} w_{|\mathcal{X}|}\mathbb{I}_{\{\mathcal{P}'\subseteq\mathcal{X}\}}$. Now, we will count the number of times the combination $(r_1, \dots, r_u)$ appears in the l.h.s. and weight it by $w_{|\mathcal{X}|}$ to compute the coefficient corresponding to $\mathcal{P}$ in the l.h.s. Trivially, if $|\mathcal{X}| < u$, then the combination will not appear (since there will not be enough number of unique players). For $|\mathcal{X}| \in \{u, \dots, |\mathbb{P}| - 1\}$, the contribution to the combination $(r_1, \dots, r_u)$ equals

$$\underbrace{w_{|\mathcal{X}|}}_{\text{weight}} \times \underbrace{\binom{|\mathbb{P}| - 1 - u}{|\mathcal{X}| - u}}_{\#\text{ of times the combination appears in the summation over } \mathcal{X} \text{ for a fixed } |\mathcal{X}|}.$$

Summing over $|\mathcal{X}|$, the total contribution (which is the coefficient we are interested in) equals

$$\sum_{|\mathcal{X}|=u}^{|\mathbb{P}|-1} w_{|\mathcal{X}|} \times \binom{|\mathbb{P}| - 1 - u}{|\mathcal{X}| - u} = \frac{1}{|\mathbb{P}| - u}\sum_{|\mathcal{X}|=u}^{|\mathbb{P}|-1}\prod_{i=0}^{u-1}\frac{|\mathcal{X}| - i}{|\mathbb{P}| - i} = \frac{1}{u+1},$$

where the last equality follows Lemma A.2. This agrees with the r.h.s. and completes our proof. ◻

Lemma A.2. *Let k_1 be a non-negative integer and k_2 be a positive integer such that $k_2 \geq k_1 + 1$. Then,*

$$\frac{1}{k_2 - k_1}\sum_{j=k_1}^{k_2-1}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2-i} = \frac{1}{k_1+1}.$$

Proof. We will prove this via induction on k_2. For the base case, consider $k_2 = k_1 + 1$. The l.h.s. equals

$$\begin{aligned}
\frac{1}{k_1 + 1 - k_1}\sum_{j=k_1}^{k_1+1-1}\prod_{i=0}^{k_1-1}\frac{j-i}{k_1+1-i} &= \prod_{i=0}^{k_1-1}\frac{k_1-i}{k_1+1-i} \\
&= \frac{1}{k_1+1},
\end{aligned}$$

which is what we wanted. For the induction step, we show that if the statement is true for some $k_2 \geq k_1 + 1$, then it is also true for $k_2 + 1$, that is,

$$\frac{1}{k_2 + 1 - k_1}\sum_{j=k_1}^{k_2}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2+1-i} = \frac{1}{k_1+1}.$$

The l.h.s. of the above equation equals

$$\begin{aligned}
&\frac{1}{k_2 + 1 - k_1}\sum_{j=k_1}^{k_2}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2+1-i} \\
&= \frac{1}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2+1-i} + \prod_{i=0}^{k_1-1}\frac{k_2-i}{k_2+1-i}\right\} \\
&= \frac{1}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2+1-i} + \frac{k_2-k_1+1}{k_2+1}\right\} \\
&= \frac{1}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\prod_{i=0}^{k_1-1}\left(\frac{j-i}{k_2-i}\times\frac{k_2-i}{k_2+1-i}\right) + \frac{k_2-k_1+1}{k_2+1}\right\} \\
&= \frac{1}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\left[\left(\prod_{i=0}^{k_1-1}\frac{j-i}{k_2-i}\right)\left(\prod_{i'=0}^{k_1-1}\frac{k_2-i'}{k_2+1-i'}\right)\right] + \frac{k_2-k_1+1}{k_2+1}\right\} \\
&= \frac{1}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\left[\left(\prod_{i=0}^{k_1-1}\frac{j-i}{k_2-i}\right)\left(\frac{k_2-k_1+1}{k_2+1}\right)\right] + \frac{k_2-k_1+1}{k_2+1}\right\} \\
&= \frac{\frac{k_2-k_1+1}{k_2+1}}{k_2+1-k_1}\left\{\sum_{j=k_1}^{k_2-1}\prod_{i=0}^{k_1-1}\frac{j-i}{k_2-i} + 1\right\} = \frac{1}{k_2+1}\left\{\frac{k_2-k_1}{k_1+1} + 1\right\} = \frac{1}{k_1+1},
\end{aligned}$$

where the second-last equality is true due to the induction hypothesis. This completes the proof. ◻

REFERENCES

[1] Vibhanshu Abhishek, Stylianos Despotakis, and R Ravi. 2017. Multi-Channel Attribution: The Blind Spot of Online Advertising. Working paper, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959778 (2017).

[2] Vibhanshu Abhishek, Peter Fader, and Kartik Hosanagar. 2012. Media exposure through the funnel: A model of multi-stage attribution. Working paper, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2158421 (2012).

[3] Eva Anderl, Ingo Becker, Florian Von Wangenheim, and Jan Hendrik Schumann. 2016. Mapping the customer journey: Lessons learned from graph-based online attribution modeling. *International Journal of Research in Marketing* 33, 3 (2016), 457–474.

[4] Sai Kumar Arava, Chen Dong, Zhenyu Yan, Abhishek Pani, et al. 2018. Deep Neural Net with Attention for Multi-channel Multi-touch Attribution. Working paper, available at: https://arxiv.org/abs/1809.02230 (2018).

[5] Nikolay Archak, Vahab S Mirrokni, and S Muthukrishnan. 2010. Mining advertiser-specific user behavior using adfactors. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 31–40.

[6] Joan Arensman and Wilfred Yeung. 2016. Move beyond last click attribution in AdWords. Retrieved Feb 8, 2018 from https://adwords.googleblog.com/2016/05/move-beyond-last-click-attribution.html

[7] Konstantin Avrachenkov, Laura Cottatellucci, and Lorenzo Maggi. 2012. Confidence intervals for Shapley value in Markovian dynamic games. Technical report (Eurecom), available at: http://www.eurecom.fr/publication/3602 (2012).

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

[9] Alexander Bleier and Maik Eisenbeiss. 2015. Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science* 34, 5 (2015), 669–688.

[10] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36, 5 (2009), 1726–1730.

[11] Hana Chockler and Joseph Y Halpern. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* 22 (2004), 93–115.

[12] Hana Choi, Carl Mela, Santiago Balseiro, and Adam Leary. 2017. Online Display Advertising Markets: A Literature Review and Future Directions. Working paper, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070706 (2017).

[13] John David Collins, Edward Jonathan Hall, Ned Hall, and Laurie Ann Paul. 2004. *Causation and counterfactuals*. MIT Press.

[14] Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. 2012. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*. ACM, 7.

[15] Peter J Danaher and Harald J van Heerde. 2018. Delusion in Attribution: Caveats in Using Attribution for Multimedia Budget Allocation. *Journal of Marketing Research* 55, 5 (2018), 667–685.

[16] Ellery Eells. 1991. *Probabilistic causality*. Vol. 1. Cambridge University Press.

[17] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. 2008. A linear approximation method for the Shapley value. *Artificial Intelligence* 172, 14 (2008), 1673–1699.

[18] Charles Miller Grinstead and James Laurie Snell. 2012. *Introduction to probability*. American Mathematical Society.

[19] Shelly Hagan. 2018. Digital economy has been growing at triple the pace of U.S. GDP. Retrieved Oct 11, 2018 from https://www.bloomberg.com/news/articles/2018-03-15/digital-economy-has-been-growing-at-triple-the-pace-of-u-s-gdp

[20] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science* 56, 4 (2005), 843–887.

[21] Christopher Hitchcock. 1997. Probabilistic causation. (1997). Retrieved Feb 1, 2019 from https://plato.stanford.edu/entries/causation-probabilistic/

[22] David Hume. 2003. *A treatise of human nature*. Courier Corporation.

[23] Marketing Science Institute. 2016. Research Priorities 2016-2018. Retrieved Feb 8, 2018 from https://www.msi.org/uploads/articles/MSI_RP16-18.pdf

[24] Wendi Ji and Xiaoling Wang. 2017. Additional Multi-Touch Attribution for Online Advertising. In *AAAI*. 1360–1366.

[25] Wendi Ji, Xiaoling Wang, and Dell Zhang. 2016. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1373–1382.

[26] Patrick Jordan, Mohammad Mahdian, Sergei Vassilvitskii, and Erik Vee. 2011. The multiple attribution problem in pay-per-conversion advertising. In *International Symposium on Algorithmic Game Theory*. Springer, 31–43.

[27] Lukáš Kakalejčík, Jozef Bucko, Paulo AA Resende, and Martina Ferencova. 2018. Multichannel Marketing Attribution Using Markov Chains. *Journal of Applied Management and Investments* 7, 1 (2018), 49–60.

[28] PK Kannan, Werner Reinartz, and Peter C Verhoef. 2016. The path to purchase and attribution modeling: Introduction to special section. *International Journal of Research in Marketing* 33, 3 (2016), 449–456.

[29] Hongshuang Li and PK Kannan. 2014. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research* 51, 1 (2014), 40–56.

[30] Yiyi Li, Ying Xie, and Eric Zheng. 2017. Modeling Multi-Channel Advertising Attribution Across Competitors. Working paper, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3047981 (2017).

[31] David Liben-Nowell, Alexa Sharp, Tom Wexler, and Kevin Woods. 2012. Computing shapley value in supermodular coalitional games. In *International Computing and Combinatorics Conference*. Springer, 568–579.

[32] Stephen C Littlechild and Guillermo Owen. 1973. A simple expression for the Shapley value in a special case. *Management Science* 20, 3 (1973), 370–372.

[33] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. 2013. Bounding the estimation error of sampling-based Shapley value approximation. Working paper, available at: https://arxiv.org/abs/1306.4265 (2013).

[34] Tomasz P Michalak, Karthik V Aadithya, Piotr L Szczepanski, Balaraman Ravindran, and Nicholas R Jennings. 2013. Efficient computation of the Shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research* 46 (2013), 607–650.

[35] Stephen L Morgan and Christopher Winship. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

[36] Guillermo Owen. 1972. Multilinear extensions of games. *Management Science* 18, 5-part-2 (1972), 64–79.

[37] Judea Pearl. 2009. *Causality*. Cambridge University Press.

[38] Colin Priest. 2017. Multichannel Marketing Attribution with DataRobot. (2017). Retrieved Oct 11, 2018 from https://www.datarobot.com/resource/mktattribution/

[39] Quantcast. 2013. Beyond Last Touch: Understanding Campaign Effectiveness. (2013). Retrieved Oct 11, 2018 from http://info.quantcast.com/rs/quantcast/images/Quantcast%20White%20Paper_Beyond%20Last%20Touch.pdf

[40] Quantcast. 2016. Guide to marketing attribution. (2016). Retrieved Oct 11, 2018 from https://www.iabuk.com/sites/default/files/case-study-docs/Quantcast%20-%20Guide%20to%20Marketing%20Attribution%20-%20EN%20-%202016%20-%20print.pdf

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

[42] Xuhui Shao and Lexin Li. 2011. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 258–264.

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

[44] Sarah Sluis. 2018. Digital ad market soars to $88 billion, Facebook and Google contribute 90% of growth. Retrieved Oct 11, 2018 from https://adexchanger.com/online-advertising/digital-ad-market-soars-to-88-billion-facebook-and-google-contribute-90-of-growth/

[45] Econsultancy Digital Marketers United. 2012. Quarterly Digital Intelligence Briefing: Making Sense of Marketing Attribution (in association with Adobe). (2012). Retrieved Oct 11, 2018 from http://success.adobe.com/assets/en/downloads/whitepaper/Adobe-Quarterly-Digital-Intelligence-Briefing-Digital-Trends-for-2013.pdf

[46] Lizhen Xu, Jason A Duan, and Andrew Whinston. 2014. Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science* 60, 6 (2014), 1392–1412.

[47] Ya Zhang, Yi Wei, and Jianbiao Ren. 2014. Multi-touch attribution in online advertising with survival theory. In *2014 IEEE International Conference on Data Mining (ICDM)*. IEEE, 687–696.

[48] Kaifeng Zhao, Seyed Hanif Mahboobi, and Saeed R Bagheri. 2018. Revenue-based attribution modeling for online advertising. *International Journal of Market Research* 61, 2 (2018), 195–209.