Multi-Touch Attribution in Online Advertising with Survival Theory

Ya Zhang* Yi Wei* Jianbiao Ren†
*Shanghai Key Laboratory of Multimedia Processing and Transmissions
†Antai College of Economics & Management
Shanghai Jiao Tong University, Shanghai, China
E-mail: {ya_zhang, ywei, jbren}@sjtu.edu.cn

***Abstract*—Multi-touch attribution, which allows distributing the credit to all related advertisements based on their corresponding contributions, has recently become an important research topic in digital advertising. Traditionally, rule-based attribution models have been used in practice. The drawback of such rule-based models lies in the fact that the rules are not derived form the data but only based on simple intuition. With the ever enhanced capability to tracking advertisement and users' interaction with the advertisement, data-driven multi-touch attribution models, which attempt to infer the contribution from user interaction data, become an important research direction. We here propose a new data-driven attribution model based on survival theory. By adopting a probabilistic framework, one key advantage of the proposed model is that it is able to remove the presentation biases inherit to most of the other attribution models. In addition to model the attribution, the proposed model is also able to predict user's 'conversion' probability. We validate the proposed method with a real-world data set obtained from a operational commercial advertising monitoring company. Experiment results have shown that the proposed method is quite promising in both conversion prediction and attribution.**

***Keywords*-Multi-touch attribution, Survival theory, Online Advertising**

I. Introduction

Over the past decade, digital advertising has become increasingly important. It has been reported that Internet ad spending has for the first time surpassed broadcast TV in 2013[1]. The driven force of such rapid growth of digital advertising comes from the following two aspects. First, increased web usages allows the Internet to better influence consumers on their purchase decisions. Secondly, digital advertising, with its ability to tracking users' interactions with the advertisements, enables more targeted delivery and hence more effective advertisement.

Digital advertising is available in a number of formats, including but not limited to textual ads, banner ads, rich media ads, and social media pages. The different forms of advertisements are delivered through multiple media channels, such as search, display, social, mobile and video. An advertising campaign is generally composed of a coordinated series of linked advertisements in multiple combinations of formats and delivery channels. To improve the return on investment (ROI) of advertising, how to allocate budget among different ad formats and ad channels becomes one of the most essential problem in advertising. Since the infancy of advertisement, many efforts have been devoted to study the effectiveness of advertising [6], [9], [11], [2]. However, most the existing studies either perform their analysis based on small scale user studies or ignored the interaction among different advertisement channels.

Typically, an individual user is exposed to multiple advertisement impressions delivered through multiple advertising channels. Attribution is to understand the contribution of different advertisements in driving users to desired actions such as clicking or making a purchase. To measure the effectiveness of different advertisements and hence optimize advertising campaigns, attribution has been recognized as one of the most critical problems in digital advertising. *Last Touch Attribution*, one of the earliest attribution models, is to give all the credit to the last advertisement a user sees before a conversion. It is widely adopted in practice and considered as standard attribution model in most web analytics tools. However, despite of its simplicity, one significant disadvantage of the last touch attribution is that it only recognizes the contribution of one single advertisement impression for any conversion and cannot credit the advertisement presented before the 'last touch'. In fact, conversion often is due to the cumulative effect of a cascade of advertisement. Simply attributing the credit to last touch may over-weight the contribution of some particular types of advertisements, such as Search advertisement, which is initiated by user queries. In fact, very often, the querying could be due to a previously viewed advertisements.

Multi-touch attribution, which allows distributing the credit to all related advertisements based on their corresponding contributions, was recently introduced and become an important research topic in digital advertising. Several rule-based attribution models have been proposed, including last-touch model, first-touch model, linear model, and time decay model, where last-touch model and first-touch model may be viewed as special cases of multi-touch attribution models. See Figure 1 for an illustration of different attribution models. The drawback of the above rule-based models lies in the fact that the rules are derived from some simple intuition and may not fit the reality well. With the ever

[1] http://mashable.com/2014/04/10/mobile-surge-internet-beats-tv/

1550-4786/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDM.2014.130

IEEE computer society

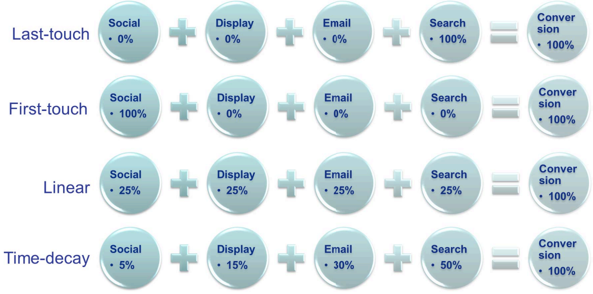


Figure 1. Illustration of rule-based methods.

enhanced capability to tracking advertisement placement and users' interaction with the advertisements, a few data-driven multi-touch attribution models [8], [3], which attempt to infer the contribution from real user interaction data, are recently proposed. However, most of the existing data-driven attribution methods suffer from the following drawbacks: 1) ignoring the presentation biases introduced by advertising placement, and 2) focusing solely on the modeling of the attribution and lacks of a solid conversion prediction model.

In this paper, we propose a new data-driven model, additive hazard model, for multi-touch attribution. In particular, the proposed model considers not only the differences in the impact strength of different advertising channels but also the variations of their time-decaying speed. Each channel is characterized by a set of two parameters, the maximum strength of its impact and the time-decaying speed of its impact. We model the time-dependent contribution of an advertisement channel using a hazard function with a set of additive exponential kernel functions, each of which is assumed to reflect the dynamics of the influences of an advertisement channel on user conversion. Based on survival theory, we further model the conversions as a user's 'death' after a cascades of advertisements. We denote the proposed model as ADDITIVEHAZARD model thereafter. The ADDITIVEHAZARD model is fitted by maximizing the log-likelihood function in an iterative manner. The major contributions of this paper are as follows. First, in addition to modeling the attribution, the proposed model is able to predict user's 'conversion' probability. Secondly, by modeling the impact of advertisement strength and time-decaying speed, we attempt to remove the presentation biases introduced by advertising placement. We validate the proposed method with a real-world data set obtained from an operational commercial advertisement monitor company. Experiment results have shown that the proposed method is quite promising in both conversion prediction and attribution modeling.

The rest of this paper is structured as follows: Section II briefly reviews the related work on multi-touch attribution. In Section III, we introduce basic concepts of survival models. The proposed Additive Hazard model for multi-touch attribution is presented in Section IV. The distributed implementation using MAPREDUCE is presented in Section . Section VI and Section VII present the experiments and results on synthesis data sets and real-world data sets, respectively. Finally, we conclude the paper and propose for future work in Section VIII.

II. RELATED WORK

The multi-touch attribution problem is generally defined as the assignment of conversion credit when multiple advertising channels reach a given online user. The most commonly used attribution methodology in the industry is generally to assign full conversion credit to a single advertising channel, typically the last advertising channel to a conversion ('last-touch' attribution) or the first advertising channel ('first-touch' attribution). However, the above models rely on the rules which are not derived form the data but only from simple intuition, which might lead to a biased estimate.

Some recent research papers have been devoted to study the conversion attribution problem with data-driven approaches. Shao and Li [8] propose two models to solve the multi channel attribution problem. The first one is a bagged logistic regression method. They combine the commonly used logistic regression, which is simple and easy to interpret, and the bagging idea, which is to help reduce the estimation variability due to the highly correlated covariants. The second one is a probabilistic model based on a combination of first and second-order conditional probabilities. For a given data set, they compute the empirical probability of the main factors

$$P(y|x_i) = \frac{N_p x_i}{N_p x_i + N_n x_i}$$

and the pair-wise conditional probabilities

$$P(y|x_i, x_j) = \frac{N_p x_i, x_j}{N_p x_i, x_j + N_n x_i, x_j}, i \neq j,$$

where y is a binary outcome variable denoting a conversion event, and $x_i, i = 1, ..., p$ denote p different advertising channels. $N_p(x_i)$ and $N_n(x_i)$ denote the number of positive or negative users exposed to channel i, respectively, and $N_p(x_i, x_j)$ and $N_n(x_i, x_j)$ denote the number of positive or negative users exposed to both channels i and j.

Xe *et al.* [10] develop a stochastic model for online purchasing and advertisement clicking that incorporates mutually exciting point processes with individual heterogeneity in a Bayesian hierarchical modeling framework. The mutually exciting point process is a multivariate stochastic process in which different types of advertisement clicks and purchases are modeled as different types of random points in continuous time. The occurrence of an earlier point affects the probability of occurrence of later points of all types so that the exciting effects among all advertisement clicks are well captured. As a result, the intensities of the point process, which can be interpreted as the instant probabilities of point occurrence, depend on the previous history of the process. They applied Bayesian inference using Markov chain Monte Carlo (MCMC) method to a mutually exciting point process model, which enables them to fit a more complex hierarchical model with random effects in correlated stochastic processes. They find that the commonly used measure of conversion rate is biased in favor of search advertisements by overemphasizing the "last click" effects and underestimates the effectiveness of display advertisements the most severely.

Dalessandro *et al.* [3] first propose the following properties a good attribution system should have-fariness, data-driven and interpretability. And then they formulate multi-touch attribution as a causal estimation problem. They present a causal framework for evaluating multi-touch attribution. They define parameters of interest that directly measure the additive marginal lift of each ad within the ad campaign and argue that this lift represents the value created by targeted and thus attribution credit should be directly derived from this measure. The basis of each attribution parameter is a counterfactual framework that under strict assumptions about the data can be interpreted as causal parameters. They discuss the practical limitations of this fully casual method and then define an approximate attribution measure that can be recast through the lens of channel importance estimation.

Abhishek *et al.* [1] use a HMM to capture the user's deliberation process and his movement down the conversion funnel as a result of the different ad exposures he experiences. In accordance with the conversion funnel, they construct an HMM with four states where the four states are "dormant", "awareness", "consideration" and "conversion". The consumer model is used to generate a new attribution technique which might be useful for future research in analyzing the impact of advertising on consumer behavior.

III. Survival Models

In this section, we briefly introduce basic concepts of survival models [7], based on which we build our own attribution model. Let T be a non-negative continuous variable representing the waiting time until the occurrence of an event, which is a conversion in our context. For simplicity, we will adopt the terminology of survival analysis, referring to the conversion of a user as 'death' and to the waiting time as 'survival' time.

A. The Survival Function

The survival function is defined as:

$$S(t) = Pr(T > t), \tag{1}$$

where t is some time and Pr stands for probability. The survival function is the probability that the time of 'death' is later than some specified time t. And then the cumulative distribution function $F(t)$ is defined as:

$$F(t) = Pr(T \leq t) = 1 - S(t). \tag{2}$$

If $F(t)$ is differentiable, the probability density function (p.d.f) $f(t)$ is :

$$f(t) = \frac{d}{dt}F(t). \tag{3}$$

B. The Hazard Function

Formally, we define the hazard function (or instantaneous rate of occurrence of the event) as:

$$\lambda(t) \quad = \quad \lim_{dt\to 0} \frac{Pr(t \leq T \leq t + dt | T > t)}{dt} \tag{4}$$

Furthermore, we have

$$\begin{aligned}
\lambda(t) &= \lim_{dt\to 0} \frac{Pr(t \leq T \leq t + dt | T > t)}{dt} \\
&= \lim_{dt\to 0} \frac{Pr(t \leq T \leq t + dt)/Pr(T > t)}{dt} \\
&= \lim_{dt\to 0} \frac{(F(t + dt) - F(t))/S(t)}{dt} \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{S'(t)}{S(t)}.
\end{aligned} \tag{5}$$

Since $\frac{d(\log g(t))}{dt} = \frac{g'(t)}{g(t)}$, where g is a differentiable function, we have

$$\lambda(t) = -\frac{d(\log S(t))}{dt}. \tag{6}$$

Thus,

$$S(t) = \exp(-\int_0^t \lambda(u)du). \tag{7}$$

IV. AdditiveHazard Model for Online Advertising

In this section, we introduce the proposed AdditiveHazard model for simultaneously modeling conversions and attribution in online advertising based on the survival models. We denote the users as $\{1, \cdots, U\}$ and the advertising channels as $\{1, \cdots, n\}$.The obeseration is a set of cascading behaviors of the users $\{c_1, \cdots, c_U\}$, each of which is in the form of $\{\{(a_i^u, t_i^u)\}_{i=1}^{l_u}, X_u, T_u\}$, where a_i^u is the advertising channel ID, t_i^u is the timestamp of impression, l_u is the length of cascade c_u and X_u is the conversion result ($X_u = 1$ means conversion). If $X_u = 1$, the last timestamp T_u is the conversion time. If $X_u = 0$, the last timestamp T_u is the observation time window.

We propose an additive risk model, AdditiveHazard, for user conversions in online advertising. We consider hazard function $\lambda_u(t)$ of user u to be additive on the clicking or viewing online advertising channels. The hazard function of cascade c_u can be expressed as:

$$\lambda_u(t) = \begin{cases} \sum_{t_i^u \le t} g_{a_i^u}(t - t_i^u), & t \le t_{l_u}, \\ 0, & otherwise \end{cases} \tag{8}$$

where $g_{a_i^u}(\cdot)$ is the the kernel function used to model the advertising channel a_i^u's effect on the conversion of user u. We select the exponential kernel function so that the dynamics of the influences of an advertisement on user conversion can be reflected by explicitly modeling the strength of impact β_k and its time-decaying property ω_k. Figure 2 shows a example of our model's kernel function and the resulting hazard function, from which we can see that the hazard function is the sum of kernel functions triggered by corresponding advertising channels. The hazard function of cascade c_u with the exponential kernel can then be expressed as:

$$\lambda_u(t) = \begin{cases} \sum_{t_i^u \le t} \beta_{a_i^u} \omega_{a_i^u} e^{(-\omega_{a_i^u}(t - t_i^u))}, & t \le t_{l_u}, \\ 0, & otherwise \end{cases} \tag{9}$$

where $\beta_{a_i^u} \omega_{a_i^u} \exp(-\omega_{a_i^u}(t - t_i^u))$ is the kernel function we choose and it decreases as time decays, $\beta_{a_i^u}$ stands for the strength of the effect triggered by advertising channel a_i^u and $\omega_{a_i^u}$ controls the speed of the time-decaying effect.

We collect the parameters into matrix-vector forms, $\boldsymbol{\beta} = (\beta_i)$ for the strength coefficients and $\boldsymbol{\omega} = (\omega_i)$ for the time-decaying coefficients. We use $\boldsymbol{\beta} \ge 0$ and $\boldsymbol{\omega} \ge 0$ to indicate that we require the matrices to be entry-wise nonnegative. The log-likelihood function of the parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\omega}\}$ can be expressed as follows.

$$\begin{aligned} \mathcal{L}_u(\Theta) &= \log((S(T_u)\lambda(T_u))^{X_u} S(T_u)^{1-X_u}) \\ &= X_u\big(\log S(T_u) + \log \lambda(T_u)\big) \\ &\quad +(1 - X_u)\log S(T_u) \\ &= X_u \log \lambda(T_u) + \log S(T_u) \\ &= \sum_{X_u=1} \log(\sum_i \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(T_u - t_i^u))) \\ &\quad - \int_0^{T_u} \sum_{t_i^u \le t} \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(t - t_i^u))dt \\ &= \sum_{X_u=1} \log(\sum_i \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(T_u - t_i^u))) \\ &\quad - \sum_i \int_{t_i^u}^{T_u} \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(t - t_i^u))dt \\ &= \sum_{X_u=1} \log(\sum_i \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(T_u - t_i^u))) \\ &\quad - \sum_i \beta_{a_i^u}(1 - \exp(-\omega_{a_i^u}(T_u - t_i^u))) \end{aligned} \tag{10}$$

In the log-likelihood function, the first part $X_u\big(\log S(T_u) + \log \lambda(T_u)\big)$ stands for the probability of user who purchased, the second part $(1 - X_u)\log S(T_u)$ stands for the probability of user who did not.

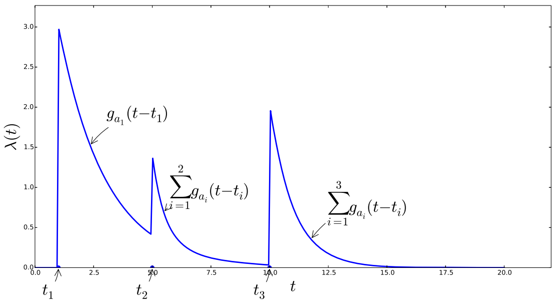


Figure 2. A simple example of our model's hazard function.

Thus, the inference problem is:

$$\begin{aligned} \underset{\Theta}{\text{maximize}} \quad & \mathcal{L}(\Theta) \\ \text{subject to} \quad & \beta_i \ge 0, \ i = 1, \ldots, n, \\ & \omega_i \ge 0, \ i = 1, \ldots, n. \end{aligned} \tag{11}$$

By maximizing the likelihood function, we can obtain the estimates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

It turns out that the above objective function can be optimized efficiently by the MM(minorize maximize) algorithm [5] in a iterative manner. In particular, we construct a lower-bound $Q(\Theta|\Theta^{(t)})$ at current estimation $\Theta^{(t)}$ as follows:

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) = &\sum_{X_u=1} \sum_i p_i^u \log \frac{\beta_{a_i^u}\omega_{a_i^u} e^{(-\omega_{a_i^u}(T_u - t_i^u))}}{p_i^u} \\ &- \sum_i \beta_{a_i^u}(1 - \exp(-\omega_{a_i^u}(T_u - t_i^u))), \end{aligned} \tag{12}$$

where p_i^u is defined as follows:

$$p_i^u = \begin{cases} \frac{\beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(T_u - t_i^u))}{\sum_{i=1}^{l_u} \beta_{a_i^u}\omega_{a_i^u} \exp(-\omega_{a_i^u}(T_u - t_i^u))}, & X_u = 1 \\ 0, & X_u = 0 \end{cases} \tag{13}$$

The p_i^u have the nice interpretations that reveal the contribution of adversing channels for the conversion of user u. Specifically, p_i^u represents the contribution of the i-th advertising channel for the conversion user u at time T_u if $X_u = 1$.

The following two properties hold for $Q(\Theta|\Theta^{(t)})$:

$$\mathcal{L}(\Theta) \ge Q(\Theta|\Theta^{(t)}), \quad \forall \Theta \tag{14}$$

$$\mathcal{L}(\Theta^{(t)}) = Q(\Theta^{(t)}|\Theta^{(t)}) \tag{15}$$

Moreover, let $\Theta^{(t+1)} = \max_\Theta Q(\Theta|\Theta^{(t)})$, we have

$$\mathcal{L}(\Theta^{(t+1)}) \ge Q(\Theta^{(t+1)}|\Theta^{(t)}) \tag{16}$$

$$\ge Q(\Theta^{(t)}|\Theta^{(t)}) = \mathcal{L}(\Theta^{(t)}), \tag{17}$$

which shows that $\mathcal{L}$ increases monotonically during the iterations and it can be shown that the iterates converges to the local optimal $\mathcal{L}_{\text{hotness}}$ [5].

Another advantage of optimizing $Q(\Theta|\Theta^{(t)})$ is that all variables $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ can be optimized independently with closed-form solution and the non-negativity constraints are naturally taken care of.

Optimizing with respect to β_k*:* Let $\frac{\partial Q}{\partial \beta_k} = 0$, we have

$$\frac{\partial Q}{\partial \beta_k} = \frac{\sum_{u,i,X_u==1,a_i^u==k} p_i^u}{\beta_k} - \sum_{i,a_i^u==k} 1 - e^{-\omega_k^{(t)}(T_u - t_i^u)}$$

The following update equation for β_k:

$$\beta_k = \frac{\sum_{u,i,X_u==1,a_i^u==k} p_i^u}{\sum_{u,i,a_i^u==k} 1 - e^{-\omega_k^{(t)}(T_u - t_i^u)}}$$

Optimizing with respect to ω_k*:* Let $\frac{\partial Q}{\partial \omega_k} = 0$, we have

$$\frac{\partial Q}{\partial \omega_k} = \sum_{u,i,X_u==1,a_i^u==k} p_i^u\left(\frac{1}{\omega_k^u} - (T_u - t_i^u)\right) - \sum_{i,a_i^u==k} \beta_k^{(t)}(T_u - t_i^u)e^{-\omega_k^{(t)}(T_u - t_i^u)}$$

Therefore, we can update ω_k as follow:

$$\omega_k = \frac{\sum_{u,i,X_u==1,a_i^u==k} p_i^u}{\sum_{u,i,a_i^u==k} p_i^u(T_u - t_i^u) + \beta_k^{(t)}(T_u - t_i^u)e^{-\omega_k^{(t)}(T_u - t_i^u)}}$$

After fitting the ADDITIVEHAZARD model, we take the conversion probability that a user touches one single advertisement channel k as the contribution of the channel k. The AdditiveHazard models the dynamics of the influences of an advertisement on user conversion by explicitly modeling the strength of influence and its time-decaying property. Hence, in calculate the contributions of any advertisement, we need to set a pre-defined observe window T. According to the above derivation, the probability of conversion in the time window T can be formulated as:

$$P(C|\beta_k, \omega_k, T) = 1 - \exp(-\beta_k(1 - \exp(-\omega_k T))).$$

V. DISTRIBUTED IMPLEMENTATION USING MAPREDUCE

Considering the large volume of advertising data used for the multi-touch attribution, we implemented the AdditiveHarzard model using MAPREDUCE framework. The distributed architecture of inference algorithm of each iteration is shown in Figure 3, following the main idea of the well-known distributed programming model, MAPREDUCE [4]. Firstly, we distribute the training data and the inferred parameters of previous iteration into different mappers, with nearly equal amount of training data on each mapper. And then, each mapper processes the training data and give the key-value output as shown in the figure, where key= a_i^u and value= v_m as shown in Figure 3. After shuffle and sort, the output of mappers with the same key go to the same reducer. In the reducers, we combine the data with the same key and output the value of β and ω of this iteration. The inference iterates until the inferred parameters converge.

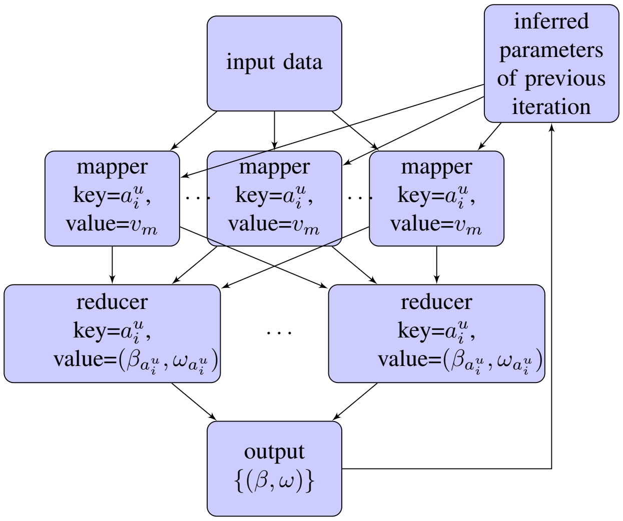

input data
inferred parameters of previous iteration
mapper key=a_i^u, value=v_m
...
mapper key=a_i^u, value=v_m
...
mapper key=a_i^u, value=v_m
reducer key=a_i^u, value=$(\beta_{a_i^u}, \omega_{a_i^u})$
...
reducer key=a_i^u, value=$(\beta_{a_i^u}, \omega_{a_i^u})$
output $\{(\beta, \omega)\}$

Figure 3. The distributed architecture of our inference algorithm. Note: The output value of mappers $v_m = (p_i^u, 1 - e^{-\omega_{a_i^u}^{(t)}(T_u - t_i^u)}, p_i^u(T_u - t_i^u) + \beta_{a_i^u}^{(t)}(T_u - t_i^u)e^{-\omega_{a_i^u}^{(t)}(T_u - t_i^u)})$.

VI. EXPERIMENTS ON SYNTHETIC DATA

To validate the proposed algorithm, we first experiment with the synthetic data. A two-step process is employed to perform the experiment: 1. Generate simulation data based on our proposed ADDITIVEHAZARD model; 2. Perform parameter inference on the data using the proposed MM alogrithm. We experiment with various parameter settings in order to understand the performance of the proposed algorithms under different settings. The parameters of the underlying ADDITIVEHAZARD model are chosen in the following way: we draw β from a uniform distribution $unif(0, 1)$ and ω from a uniform distribution $unif(1, 10)$. In the data generation step, we set the time window to 10 and number of users to 10000.

We use relative error to measure the accuracy of the parameter inference. Suppose β^* and $\hat{\beta}$ are the true and inferred parameters, respectively. The relative error is defined as follows:

$$\text{relative error} = \frac{|\beta^* - \hat{\beta}|}{\beta^*}.$$

As we can see from Table I, the relative error of all parameters is generally less than 1.4%, suggesting that the proposed ADDITIVEHARZARD method can accurately infer the parameters.

Table I
RELATIVE ERRORS OF INFERRED PARAMETERS.

β^*	ω^*	**Avg. record length**	**Conversion rate**	**Relative error of β**	**Relative error of ω**
0.463565	3.2648	1.10531	0.56305	0.00177014	0.00563106
0.294258	4.30407	1.08586	0.41912	0.000997106	0.0131392
0.423249	2.04359	1.10514	0.52856	0.000803873	0.00406164
0.758402	8.83509	1.09455	0.72663	0.00876188	0.0100482
0.391734	8.54506	1.08888	0.50772	0.00269725	0.00135432
0.579898	5.09531	1.10532	0.63764	0.00091742	0.00130336
0.97636	2.64656	1.0994	0.79586	0.00317727	0.00477636
0.171916	8.23815	1.06038	0.27889	0.00241287	0.0108729

VII. EXPERIMENTS ON REAL DATA

A. Dataset description

We use the data set openly published by Miaozhen System[2], which inludes about 380 million cookies. The data are collected from PCs and mobile devices from May 1 ,2013 to June 30, 2013 and from April 4, 2013 to June 9, 2013, respectively. Every time a user clicks on or is exposed to one of the online advertisement channels, the exact time of the event and the ID of the corresponding online advertisement channel are recorded. The information concerning users' identities is primarily determined by tracking cookies stored on their computers and encrypted in consideration of privacy. Besides that the company also provides the purchase data recording the exact time and user ID of each purchase. With all the data mentioned above, we are able to construct the historical time line of the clicks, exposures of advertisement channels and purchases for each user. We also know the type of advertisements and the website for each advertising channel.

The data contain 377,008,065 users, 10,240,200 ad clicks, 1,200,101,507 ad exposures and 4,281 conversions. There are 2605 ad channels with 46 different types and 35 websites. We present distributions of the data in Figure 4, from which we can see that the distributions of ad clicks, exposures and conversions and the the distributions of channel appearances of different types and sites all show a long-tailed pattern.

B. Experiment Setup

We get the user ID, the advertising channel ID, time, and click or exposure identification from the above data set and put them in the form of $\{\{(a_i^u, t_i^u)\}_{i=1}^{l_u}, X_u, T_u\}$ as explained in section IV. We split the whole data set into training data and testing data, each of which is 50% of the original data set. Then we fit the proposed ADDITIVEHAZARD model, last-touch, logistic regression, simple probabilistic model, and causal attribution model using the training data, and present the attribution of different advertisement channels. While it is difficult to directly validate the attribution models, we valid the models by further predicting the conversions in the test data based on the above attribution models.

Three metrics are employed to measure the performance of conversion prediction: *precision*, *recall* and *f1-score*. Suppose X_u^* and $\hat{X_u}$ are the true and inferred results of conversion ($X_u = 1$ means a conversion), respectively. The precision, recall, and f1-score are defined as follows:

$$\mathrm{P} = \frac{\sum_u (X_u^* == \hat{X_u})}{\sum_u 1}$$

$$\mathrm{R} = \frac{\sum_u (X_u^* = 1 \text{ AND } \hat{X_u} = 1)}{\sum_u X_u^* = 1}$$

$$\mathrm{F1} = \frac{2PR}{P + R}.$$

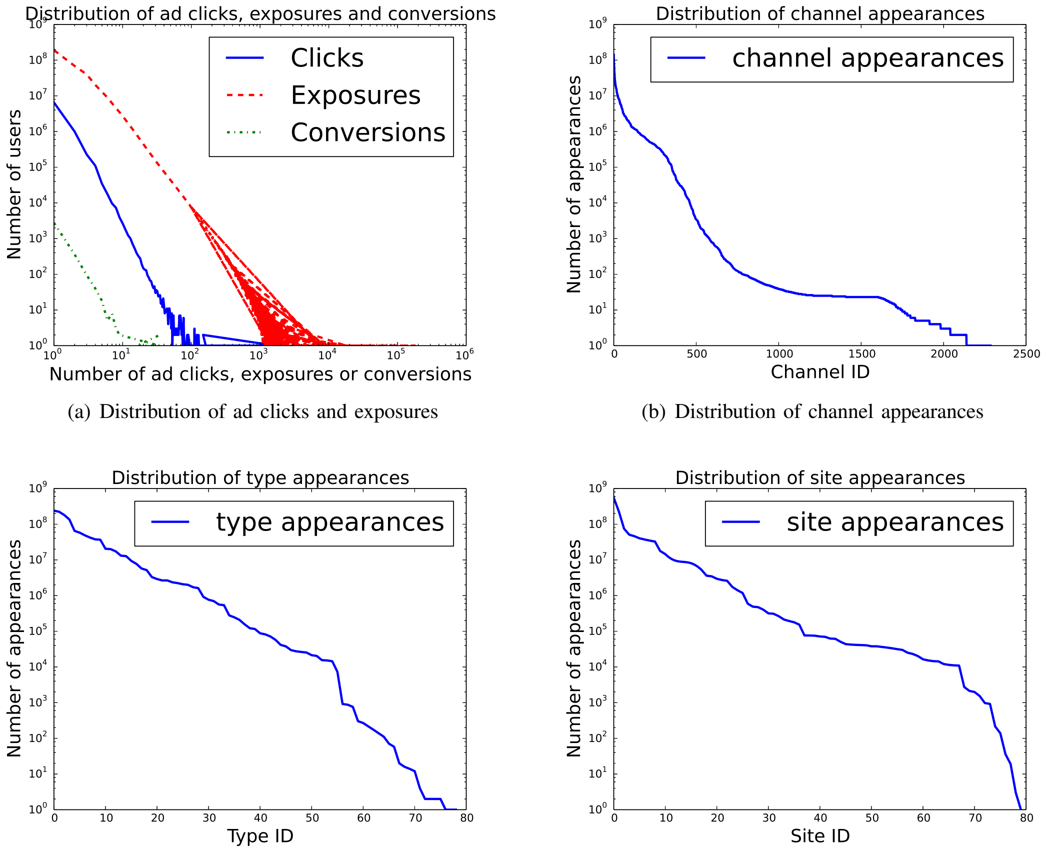
For each given advertisement cascade, we calculate its conversion probability in the time window. The top N scored users are selected as the users who will convert during the time window. By varying N, the number of users we choose, we generate the precision-recall and f1-score graph to compare the performance of different methods.

C. Interpretation of Model Parameters for ADDITIVEHAZARD

The AdditiveHazard models the dynamics of the influences of an advertisement on user conversion by explicitly modeling the strength of impact β_k and its time-decaying property ω_k. While it is impossible to evaluate the accuracy of the model fitting for β_k and ω_k, we here try to present a qualitative analysis of the channels with the highest and lowest value of β_k and ω_k.

Table II shows that 8 among the 10 channels with the largest β_k value are of the type *SEM* and on the site *SearchEngine1*, suggesting that *SearchEngine1* is a high impact web site and *SEM* is an influential type of advertising. A further analysis of the data set reveals that SEM advertisements are all Search Clicks, which is known to having high influence on user conversion. On the other hand, the 10 channels with the lowest β_k are of the type *Video(15s)* and on various video site, suggesting that video site so far is a less effective medium and *Pre-moive* ads are less influential type of advertising.

[2] http://www.miaozhen.com/BDC2013.html



(a) Distribution of ad clicks and exposures

(b) Distribution of channel appearances

(c) Distribution of type appearances

(d) Distribution of site appearances

Figure 4. Data distribution.

Table II
THE CHANNELS WITH THE HIGHEST & LOWEST IMPACT STRENGTH β_k.

Channel	Type	Site	β_k	ω_k
100234261	Column	SearchEngine1	1.69	0.03
100242639	SEM	SearchEngine1	0.51	7.93
100242089	SEM	SearchEngine1	0.30	0.0025
100291542	SEM	SearchEngine1	0.17	0.0068
100281116	SEM	SearchEngine1	0.11	0.47
100281075	SEM	SearchEngine1	0.09	0.043
100281076	SEM	SearchEngine1	0.06	0.31
100234135	Button	SearchEngine1	0.06	0.43
100275049	SEM	SearchEngine1	0.05	77.37
100242479	SEM	SearchEngine1	0.05	71.13
100242745	Pre-moive15s	VideoSite3	1.1e-6	0.0053
100256638	Back bomb	Portal1	1.1e-6	0.025
100262130	Video(15s)	VideoSite6	1.1e-6	0.10
100242709	Video(15s)	VideoSite4	1.1e-6	0.0053
100262175	Video(15s)	VideoSite5	10.0e-7	0.0053
100262158	Video(15s)	VideoSite6	9.8e-7	0.0044
100242450	iFocus	VerticalSite7	9.7e-7	0.051
100262062	Video(15s)	VideoSite4	9.2e-7	0.0057
100256703	iFocus	VerticalSite25	9.2e-7	0.10
100275296	SEM	SearchEngine1	7.8e-7	0.0054

Table III shows that 9 out of the 10 channels with the largest ω_k value are of the type *SEM* and on the site *SearchEngine1*, suggesting that the impact of the *SEM* type at the site *SearchEngine1* decays very fast. SEM is known for its high influence on user conversion. However, our results have also revealed that the impact of some of the SEM advertisements may decay fast. This could be explainable: since SEM is generally user initiated, its impact could diminish very fast if the SEM does not match the user intension well. On the other hand, the 10 channels with the lowest ω_k are of various types, such as *Promoted-Download, iFocus, SEM, and Banner* and on various portal sites and search engine sites. An interesting finding is that all these channels are associated with quite low values of β_k, suggesting that some low influential advertisement could continuously give users a low-level stimulus for conversion.

D. *Conversion Prediction*

A challenge in validating various attribution models is that there is no 'ground truth' attribution score available. So we need to find an alternative benchmark to evaluate different attribution methods. In this section, we test the five attribu-

Table III
THE CHANNELS WITH THE HIGHEST & LOWEST TIME DECAY ω_k.

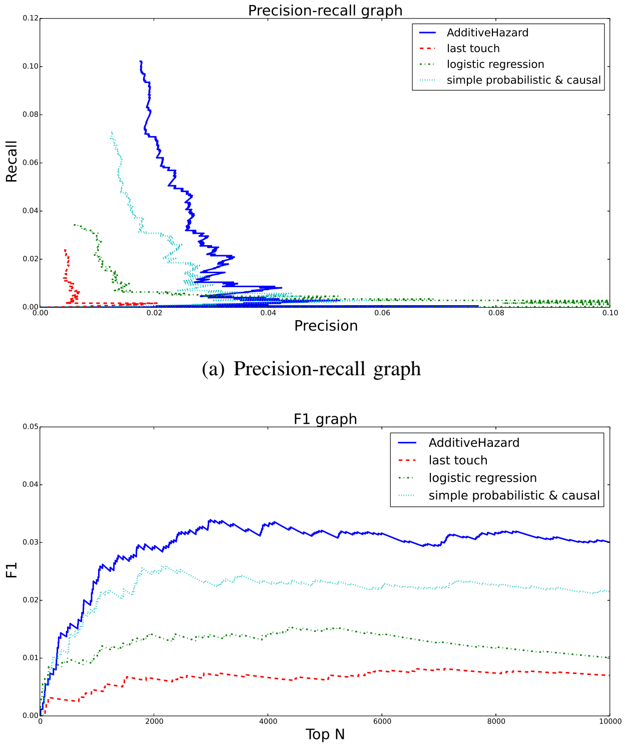
Channel	Type	Site	β_k	ω_k
100275048	SEM	SearchEngine1	0.027	278.73
100275520	SEM	SearchEngine1	0.0033	255.92
100275034	SEM	SearchEngine1	0.0029	187.81
100248636	Branded Album	SearchEngine1	0.0031	183.10
100242644	SEM	SearchEngine1	0.0033	180.44
100242334	SEM	SearchEngine1	0.0025	141.44
100275033	SEM	SearchEngine1	0.0022	128.95
100242474	SEM	SearchEngine1	0.0080	115.24
100275238	SEM	SearchEngine1	0.0036	115.02
100242477	SEM	SearchEngine1	0.0019	113.95
1Nx	PromotedDownload	VerticalSite11	1.5e-6	0.0013
100234134	Column Area	SearchEngine1	1.7e-5	9.5e-4
100281794	iFocus	Portal1	1.9e-5	7.9e-4
100281786	iFocus	Portal5	1.4e-5	7.3e-4
100281091	SEM	SearchEngine1	4.1e-5	6.4e-4
100281089	SEM	SearchEngine1	3.3e-5	5.6e-4
100281056	SEM	SearchEngine1	2.2e-5	5.2e-4
100281085	SEM	SearchEngine1	4.0e-5	4.1e-4
100281341	Banner	Portal1	6.0e-5	4.1e-4
100281341	Banner	Portal1	4.8e-5	3.8e-4

tion models with conversion prediction. It is straight-forward to perform conversion prediction with ADDITIVEHAZARD model, logistic regression model and last-touch model. For simple probabilistic model and causal attribution model, we take advantage of the simulation method in Dalessandro *et al.*'s paper [3] for conversion prediction. We use their conversion-generation formula to calculate the probability of conversion of a user:

$$P(Y) = (1 - \prod_{k}^{K}(1 - P(Y|C_k))) \times \delta^{\sum I(C_k)}$$

The first term in brackets represents the probability of conversion assuming zero interaction effects. The second term accounts for the marginally decreasing effect of each ad. We also choose $\delta = 0.95$ as they did. Notice that the formula only uses the first order conditional probabilities. This formula is used for both simple probabilistic model and causal attribution model. Therefore, the two models give the same results in terms of conversion prediction and we treat them as one single method for this purpose.

We then compare the performance of conversion prediction of the following four methods, ADDITIVEHAZARD model, last-touch model, logistic regression model, and simple probabilistic model & causal model. As we can see from Figure 5, the ADDITIVEHAZARD model clearly outperforms the other three models. Considering the purchases are extremely sparse in the real world data set (i.e., only 4,281 purchases for over 1.2 billion ad exposures), it is quite challenging to accurately predict the purchases. The ADDITIVEHAZARD model achieve nearly the F1-score of 0.035 and has much more superior performance than the other three models in terms of precision, recall, and F1 score. The next best model is the simple probabilistic model & causal model, followed by the logistic regression model. And no surprise, the last touch model performs the worse among the four models. The simple probabilistic model & causal model are more close to the ADDITIVEHAZARD model in terms of model formulation, which matches with our experimental results that in terms of precision, recall, and F1-score in conversion prediction, the two models are more similar to the ADDITIVEHAZARD model. Logistic regression model does not have a probabilistic interpretation and is more close to last touch in the attribution formulation. It is observed that the logistic regression model performs more similar to the last touch models.



(a) Precision-recall graph

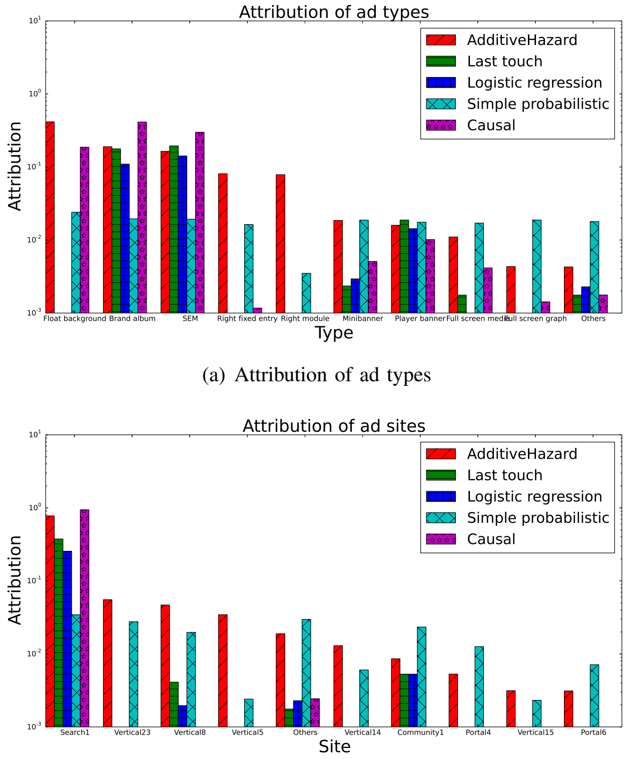
(b) F1-score graph

Figure 5. Results of conversion prediction for the real world data set by the four different attribution models.

E. Attribution Analysis

We next present the attribution analysis of 5 different methods, including the proposed ADDITIVEHAZARD model, last-touch model, logistic regression model [8], simple probabilistic model [8] and causal attribution model [3]. For the ADDITIVEHAZARD model, the dynamics of the influences of an advertisement on user conversion are modeled with the strength of influence and its time-decaying property. Hence, in perform the attribution, we need to set a predefined observe window T. Here we set it empirically to

30 days. Because it is difficult to interpret the attribution at individual channels, we here aggregate the attribution of individual channels based on the corresponding ad types or ad sites. The results are summarized in Figure 6. To facilitate the interpretation of the attribution, we also plot the distribution of distance to conversion for different ad types and ad sites in Figure 7.



(a) Attribution of ad types

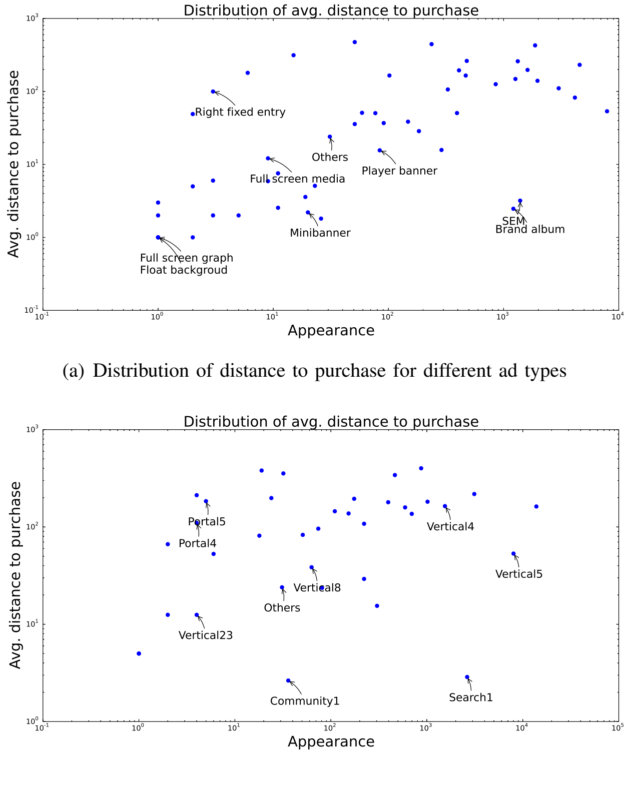
(b) Attribution of ad sites

Figure 6. Attribution analysis on real data set.

As can be seen from Figure 6 and Figure 7, the ADDITIVEHARZARD model, the simple probabilistic model, and the causal attribution model have more similar attribution, while the logistic regression model and the last touch model have more similar attribution, which matches our previous analysis that the ADDITIVEHARZARD model is more close to the simple probabilistic model and the causal attribution model, and the logistic regression model is more close to last touch model in attribution modeling.

All the models tends to give more credit to the ad types and web sites which are closer to the conversions on average and appear a large number of times in the users' ad viewing/clicking history. From Figure 6, we can find that the ad types *Brand album* and *SEM* and the site *Search1* are given most credit in the attribution. Figure 7 shows that *Brand album*, *SEM* and *Search1* all have an average distance of nearly one to conversion.

The type *Float background* is given high attribution score by the ADDITIVEHARZARD model, the simple probabilistic model, and the causal attribution model, while the logistic regression model and the last touch model give it close to 0 attribution. A further investigation into the data reveals that the type *Floating background* contains only one channel 100246500. It only appears once in last touch in 1704 conversions. Thus last touch attribute and logistic regression model tend to give it low attribution score. On the other hand, the total number of appearance of this channel in the training data is only 152, among while 37 is in last position. So the ADDITIVEHARZARD model tend to give it high value for the attribution. The main reason we attribute so different from Last touch is because we focus on the inferred more objective conversion rate of the channels by removing the influence of the presentation biases in the training data.



(a) Distribution of distance to purchase for different ad types

(b) Distribution of distance to purchase for different ad sites

Figure 7. Distribution of distance to purchase of different ad types and ad sites.

VIII. CONCLUSIONS AND FUTURE WORK

We have presented a fully data-driven model called additive hazard model based on survival theory for the multi-channel attribution problem in online advertising. In particular, the proposed model not only considers the different levels of impact of different advertising channels but also takes time-decaying effect into account. The proposed model is fitted by optimizing the likelihood function in an iterative manner. When experimented with the synthetic data and real-world data, the proposed method makes a good inference of the parameters, in terms of relative error. For the real-world data, the proposed method typically outperforms the last-touch (default model in industry) and logistic regression model.

There are several interesting directions in the future work. Currently we fit the model at individual channel level. As Shao *et al.* [8] point out, selecting the right dimensions to model on and controlling the dimensionality are important because introducing unnecessary dimensions might introduce noise and make results difficult to interpret and higher dimensionality and cardinality would either significantly increase the amount of data needed for statistical significance or drown out the important conclusions. A possible future research direction is to investigate how to limit the set of dimensions by aggregating ads of similar behaviors. In addition, a more efficient algorithm may be explored to solve the inference problem. How to find the best kernel function for each advertising channel statistically is another interesting topic. Moreover, more factors can be taken into account to be added to our model like the mutual exciting effects among different advertising channels and so on.

References

[1] V. Abhishek, P. S. Fader, and K. Hosanagar. Media exposure through the funnel: A model of multi-stage attribution.

[2] L. F. Bright and T. Daugherty. Does customization impact advertising effectiveness? an exploratory study of consumer perceptions of advertising in customized online environments. *Journal of Marketing Communications*, 18(1):19–37, 2012.

[3] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 7. ACM, 2012.

[4] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[5] D. R. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[6] R. J. Lavidge and G. A. Steiner. A model for predictive measurements of advertising effectiveness. *Journal of marketing*, 25(6), 1961.

[7] J. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley, 2011.

[8] 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.

[9] L. C. Ueltschy and R. F. Krampf. The influence of acculturation on advertising effectiveness to the hispanic market. *Journal of Applied Business Research (JABR)*, 13(2):87–102, 2011.

[10] L. Xu, J. A. Duan, and A. B. Whinston. Path to purchase: A mutually exciting point process model for online advertising and conversion. *Available at SSRN 2149920*, 2012.

[11] G. Zenetti, T. H. Bijmolt, P. S. Leeflang, and D. Klapper. Search engine advertising effectiveness in a multimedia campaign. *International Journal of Electronic Commerce*, 18(3):7–38, 2014.