A Probabilistic Multi-Touch Attribution Model for Online Advertising

Anonymous Authors

ABSTRACT

It is an important problem in computational advertising to study the effects of different advertising channels upon user conversions, as advertisers can use the discoveries to plan or optimize advertising campaigns. In this paper, we propose a novel Probabilistic Multi-Touch Attribution (PMTA) model which takes into account not only which ads have been viewed or clicked by the user but also when each such interaction occurred. Borrowing the techniques from survival analysis, we use the Weibull distribution to describe the observed conversion delay and use the hazard rate of conversion to measure the influence of an ad exposure. It has been shown by extensive experiments on a large real-world dataset that our proposed model is superior to state-of-the-art methods in both conversion prediction and attribution analysis. Furthermore, a surprising research finding obtained from this dataset is that search ads are often not the root cause of final conversions but just the consequence of previously viewed ads.

Keywords

computational advertising, multi-touch attribution, survival analysis

1. INTRODUCTION

Internet increasingly becomes the leading advertising medium, where online users generate a tremendous amount of feedback information including clicks and conversions. The feedback data reveal the needs/preferences of users, and thus enable online advertising systems to deliver ads to those who are most likely to respond. Nowadays companies spare no effort to attract consumers to visit their websites through various advertising channels, among which display ads and search ads are two dominant types.

Recently, researchers from both academia and industry have become more and more interested in analysing the contribution of each advertising channel to user conversion which is known as the "attribution" problem. An accurate attribution model would be of great help for advertisers to interpret the effects of different advertising channels and make informed decisions to optimize their advertising campaigns (e.g., by reallocating advertising budgets).

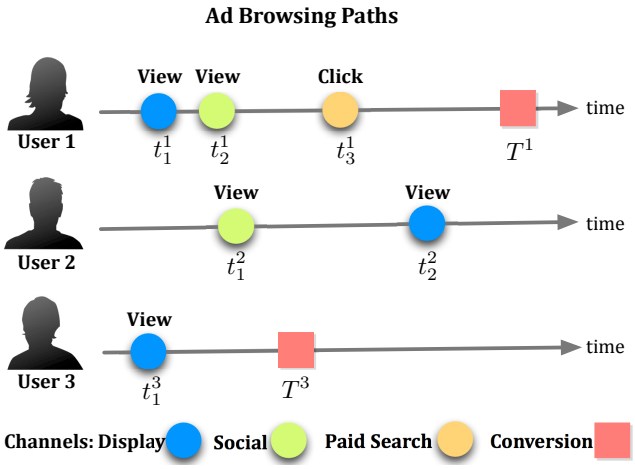


Figure 1: The possible behavioral paths in an online advertising system. Each such path consists of the chronological sequence of a user's interactions with three advertising channels: display ads, social ads, and paid search ads.

An online advertising campaign is usually launched across multiple channels such as display ads, paid search ads, social media ads, and so on. In most cases, users would have been exposed to the ads from a particular advertising campaign many times before their final conversion, as illustrated in Figure 1. Suppose that a brand `X` delivers ads through three channels: display, social and paid search: `user 1` saw `X`'s display ad at t_1^1 when browsing a webpage, and then saw `X`'s social ad at t_2^1; later, she searched for `X`'s products and clicked its paid ad link at t_3^1; finally, she made a purchase on `X`'s website at time T^1. In this case, how should we assess the contribution of those three advertising channels to that user's conversion?

A number of attribution models have been proposed and utilized in recent years. Figure 2 shows some representative ones. Most of the existing attribution models widely used in practice are rule-based, and their effectivenesses are limited by their underlying assumptions. For example, the last interaction attribution model — one of the earliest and simplest attribution models — assumes that a user's conversion is just caused by the last ad she clicked or viewed before

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

ACM ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983787

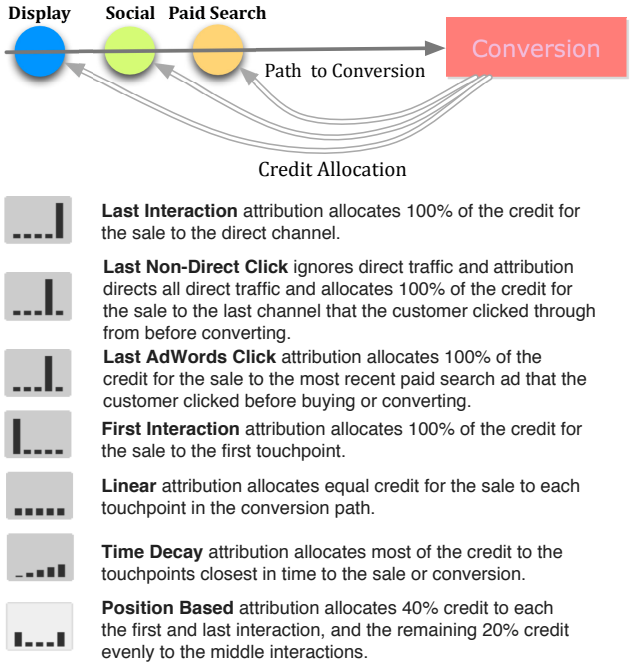


Figure 2: Examples of existing attribution models.

the conversion. It has been used as the standard attribution model in the digital marketing industry. In this model, for `user 1`, all the credits for her conversion would be completely assigned to the paid search ad she clicked in the end and the effects of all the previously viewed ads would be totally ignored. Despite its advantage of simplicity and easiness of implementation, this model obviously overestimates the contribution from the last ad. In fact, the search for `X`'s products and the subsequent click on its paid search ad — the last ad in the above example — could well be triggered by the user's previously viewed ads. A reliable attribution model should uncover the contributions from all relevant ads clicked or viewed on the user's behavioral path towards the final conversion.

Multi-Touch Attribution (MTA) aims to track how a user interacts with different advertising channels and what actions they take after each ad exposure. It has become a very hot research topic and has been studied extensively by major marketing analytics companies (e.g. Google Analytics[1], Multitouch Analytics[2], Nielsen[3]). A few data-driven models for MTA have appeared in *computational advertising* [4, 14, 16, 17, 20] (see Section 2). However, these relatively new attribution models are either totally oblivious of the temporal dimension or solely focusing on the temporal dimension. In our opinion, the ideal MTA model must take into account not only *which* ads have been viewed or clicked by the user but also *when* each such interaction occurred. First, the influence of an ad upon a user is apparently dependent on that user's interest in the product or service being advertised: if she was not interested she would never be converted. In other words, different users have different intrinsic conversion rates with respect to a particular advertising campaign, which should be captured by the model. Second, the influence of an ad upon a user is dependent on the time of its exposure to that user: she is more likely to be affected by more recent ads. Therefore, the conversion "delay" (the time interval between an ad exposure and the eventual conversion) should also be incorporated into the model.

In this paper, we propose a novel data-driven attribution model named Probabilistic Multi-Touch Attribution (PMTA) which combines the above mentioned two aspects in a consistent and coherent probabilistic framework. It is inspired by *survival analysis* [7, 8], a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms. The Weibull distribution, a commonly used lifetime distribution, is used to describe the decaying of each individual advertising channel's influence along with time. Making an analogy between the event of death and the event of conversion, we argue that the hazard rate derived from an advertising channel's survival function could be used to quantitatively measure its effect upon a user in terms of conversion.

Furthermore, we apply the proposed PMTA model to the task of conversion prediction which would provide helpful information to advertisers when they try to plan or optimize an advertising campaign using multiple advertising channels. More specifically, the PMTA model is used to predict (i) whether a user will convert and (ii) if so when she will convert, considering the combined effect of all relevant ads she has viewed or clicked. It turns out to be difficult to calculate the probability of conversion directly. One key trick here is that it would be much easier to calculate the probability of non-conversion (which happens if and only if all relevant ads so far have failed to trigger conversion) first, and then the probability of conversion is simply one minus the probability of non-conversion.

The remainder of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we explain our proposed PMTA model in detail. In Section 4, we present the experimental evaluation and discuss the results. In Section 5, we make concluding remarks.

2. RELATED WORK

In the field of computational advertising, although most existing work on conversion prediction is based on the simple last interaction attribution model, some data-driven models have recently been developed to address the problem of MTA. Shao and Li [14] have proposed a bagged logistic regression model to predict the conversion rate based on the viewed ads of a user, which is probably the first study in this area. Their approach characterizes the ad browsing path with the counts of ad exposures and estimates the credits of different advertising channels by the parameters of the trained regression model. However, the time factor has not been considered at all, and the attribution results are difficult to interpret. Dalessandro et al. [4] have formulated MTA as a causal estimation problem, and used the additive marginal lift of each ad to measure its contribution to conversion. However, their method for unbiased estimation of the causal parameters is qutite complicated and hard to implement, so the authors had to make simplistic assumptions in order to approximate their model in practice. Zhang et al. [20] have used the additive hazards model from survival analysis to estimate the temporal influence of an advertis-

[1]http://analytics.google.com
[2]http://www.multitouchanalytics.com
[3]http://www.nielsen.com

ing channel. Their decay function is made from one or more additive exponential functions. Manchanda et al. [11] have used a proportional hazards model to predict the conversion time based on the viewed ads of users. It is similar to the logistic regression method proposed by [14] except that the former aims at the prediction of conversion time while the latter aims at the prediction of conversion rate. Wooff et al. [16] used the beta distribution to model the influence of each ad, which attributes most credit to the first ad and the last ad. Xu et al. [17] proposed a model based on mutually exciting point process, which considers ad clicks and purchases as independent random events in continuous time. Chapelle [3] used exponential distribution to model the delayed feedback of clicked ads., There exist a few surveys on various models for MTA, e.g., [2, 6, 10]. One common drawback of the above models is that the differences between users' intrinsic conversion rates have been ignored.

More generally, this work is also related to the studies of user behavior based on survival analysis. For customer churn, Bolton [1] and Gonul et al. [5] used proportional hazards model to predict the probability of a customer switching to competitors. For recommendation systems, Wang and Zhang [15] used a similar method to decide the right time for recommending a product to a user. For social media, Zhang et al. [18, 19] have employed survival analysis techniques to investigate how long a Wikipedia editor will stay in the community, i.e., remain active in editing.

3. APPROACH

In this section, we briefly present the basic concepts of survival analysis which are the fundamentals of modeling the influence of an ad exposure and its decay speed, and then we carefully explain the proposed PMTA model which integrates the joint influence of all relevant ads based on the observed conversions and their delays.

3.1 Conversion Delay

As we have mentioned, survival analysis is a common approach to fine-grained modeling of the observed product lifetime in various application domains, including medicine, economics, engineering and behavior sciences [12]. Here we assume that the conversion delay T between an ad exposure and the eventual conversion, which is the lifetime in this work, follows a duration distribution with the probability density function $\varphi(t)$. The survival function $S(t)$ is defined as the cumulative probability that the conversion time is later than a specified time point t. And the hazard rate $h(t)$ presents the occurrence rate of the conversion at time point t subject to the condition that the user has not converted before t, which is defined as [9]:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t | T > t)}{\Delta t} = \frac{\varphi(t)}{S(t)} \ . \qquad (1)$$

One of the most widely used lifetime distributions in survival analysis is the Weibull distribution which is characterized by two parameters, the shape parameter α and the scale parameter λ [13]:

$$\begin{aligned} \varphi(t) &= \underbrace{\frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}}_{h(t)} \underbrace{\exp\left[-\left(\frac{t}{\lambda}\right)^{\alpha}\right]}_{S(t)} \\ &= h(t) \qquad S(t) \end{aligned} \qquad (2)$$

The factorization of $\varphi(t)$ indicates that the probability of an event occurring at time t is in fact the product of hazard rate $h(t)$ and survival probability $S(t)$.

We select the Weibull distribution to describe the duration of the observed conversion delay, and then hazard rate can reflect the influence of an ad exposure on user conversion. The shape parameter α and the scale parameter λ of the Weibull distribution determine the influence strength and its decay speed. The Weibull distribution is chosen to model conversion delay because it is a fairly versatile lifetime distribution: it can take on the characteristics of various other distributions by setting the shape parameter α appropriately, while the scale parameter λ indicates how concentrated or spread out the distribution is. As shown in Figure 3, it can be seen that the shapes of probability density function, hazard rate and survival function would take on a variety of forms according to the value of the shape parameter α.

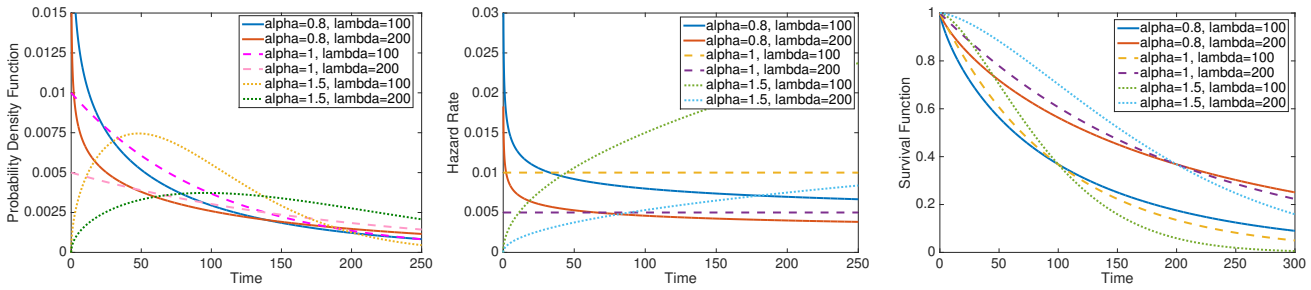
- When $\alpha < 1$, the hazard rate is a monotonic decreasing function, indicating that the occurrence rate of an event decreases over time. For conversion delays, α is usually smaller than 1, which means that the influence of an ad fades away quickly with time.
- When $\alpha = 1$, the hazard rate is a constant over time $1/\lambda$. In this situation, the Weibull distribution reduces to an exponential distribution.
- When $\alpha > 1$, the hazard rate is a monotonic increasing function, indicating that the occurrence rate of an event increases over time. The peak of the distribution will not be at $t = 0$, which is different from the situation $\alpha \leq 1$.

The advantages of the Weibull distribution over other common probability distributions for modeling lifetime duration have been proved in many practical applications [13].

3.2 Probabilistic Model

In this paper, we aim to build a probabilistic model to analyze the contribution of each ad exposure to the conversion based on the historical behavior of users.

As shown in Figure 1, an ad exposure event is defined as a user viewing or clicking an ad on an advertising channel at some time point. Before going to the details of the proposed model, let us introduce the notations used in this paper. We denote users as $\{1, ..., U\}$, and the advertising channels as $\{1, ..., K\}$. An ad browsing path b^u of user u is $\{\{a_i^u, t_i^u, x_i^u\}_{i=1}^{l_u}, Y^u, T^u, T_c^u\}$, where l_u is the length of the ad browsing path b_u, x_i^u is a set of features (including the information of viewed/clicked ads and user preferences), a_i^u is the advertising channel, t_i^u is the timestamp of impression or click, $Y_u \in \{0, 1\}$ indicates whether a conversion has already occurred and T_u is the last timestamp of the observation window. T_c^u is the timestamp of the conversion if $Y_u = 1$, and undefined otherwise. If a user does not convert in an observation window, either she will never convert or she will convert later. So there is an extra variable, $C_u \in \{0, 1\}$ indicating whether a user will eventually convert, should be incorporated into the model. Besides, $x_{c,i}^u$ and $x_{d,i}^u$ are two subsets of x_i^u, which include contextual information such as user preferences, recent impressions and clicks, etc. $x_{c,i}^u$ determines whether the conversion will occur and $x_{d,i}^u$ determines when the conversion will occur.



(a) Probability density function (b) Hazard rate (c) Survival function

Figure 3: The Weibull distribution characterized by the shape parameter α and the scale parameter λ.

First, we present the probability that a user would be converted within the observation window ($Y = 1$). It turns out to be difficult to calculate the probability of conversion directly. One key trick here is that it would be much easier to calculate the probability of non-conversion (which happens if and only if all relevant ads so far have failed to trigger conversion) first, and then the probability of conversion is simply one minus the probability of non-conversion. So the probability of user u converting at time point T_c^u would be calculated as:

$$\Pr(Y = 1|b^u) = 1 - \prod_{i=1}^{l_u} \left[1 - \Pr(Y = 1, D = d_i^u | X_c = x_{c,i}^u, X_d = x_{d,i}^u, E = e_i^u)\right] , \tag{3}$$

where $d_i^u = T_c^u - t_i^u$ represents the conversion delay of an ad exposure and $e_i^u = T^u - t_i^u$ represents the elapsed time. The product of probabilities on the right-hand side of this equation represents the probability of non-conversion, i.e., the failure of all the ads in the browsing path. In other words, the probability that the user u converts at T_c^u is the probability that at least one of the ads in the browsing path had successfully influenced the user.

Similarly, the probability that user u has not converted up to the time point T^u, i.e., all ads in the browsing path b_u have failed to trigger a conversion before T^u, can be expressed as:

$$\Pr(Y = 0|b^u) = \prod_{i=1}^{l_u} \Pr(Y = 0|X_c = x_{c,i}^u, X_d = x_{d,i}^u, E = e_i^u) , \tag{4}$$

In summary, $\Pr(Y = 1|b^u)$ represents the probability that the user converts at T_c^u and $\Pr(Y = 0|b^u)$ represents the probability that the user has not converted within the observation window (before T^u). Note that the sum of those two probabilities is not 1, because $\Pr(Y = 1|b^u)$ is the probability that the conversion occurs right at the time point T^u, but not the probability that the conversion has occurred on or before T^u. The reason of modeling $\Pr(Y = 1|b^u)$ and $\Pr(Y = 0|b^u)$ is that they are the variables directly observable and thus learnable from the log data.

Obviously if the a user u has already converted ($Y = 1$), the variable C would be observed as $C = 1$. Considering the intrinsic conversion rate and the conversion delay separately, we calculate the probability of ad exposure $\{a_i^u, t_i^u, x_i^u\}$ triggering the conversion at T_c^u as:

$$\begin{aligned} &\Pr(Y = 1, D = d_i^u | X_c = x_{c,i}^u, X_d = x_{d,i}^u, E = e_i^u) \\ =&\Pr(C = 1|X_c = x_{c,i}^u) \Pr(D = d_i^u | X_d = x_{d,i}^u, C = 1) \\ =&p(x_{c,i}^u)\varphi_k(x_{d,i}^u, d_i^u) , \end{aligned} \tag{5}$$

where $p(x_{c,i}^u)$ and $\varphi_k(x_{d,i}^u, d_i^u)$ with $k \in \{1, ..., K\}$ are both generalized linear functions to realize personality prediction. The first function $p(x_{c,i}^u)$ aims to model the (time-independent) intrinsic conversion rate by logistic regression, one of the most widely used classification models:

$$p(x_{c,i}^u) = \frac{1}{1 + \exp(-\omega_c^T x_{c,i}^u)} . \tag{6}$$

The second function $\varphi_k(x_{d,i}^u, d_i^u)$ is a Weibull distribution of the (nonnegative) conversion delay:

$$\varphi_k(x_{d,i}^u, d_i^u) = \frac{\alpha_k}{\lambda_k(x_{d,i}^u)} \left(\frac{d_i^u}{\lambda_k(x_{d,i}^u)}\right)^{\alpha_k - 1} \exp\left[-\left(\frac{d_i^u}{\lambda_k(x_{d,i}^u)}\right)^{\alpha_k}\right] . \tag{7}$$

There is a separate variable representing the conversion decay $\varphi_k(x_{d,i}^u, d_i^u)$ for each channel so as to avoid the bias caused by different media formats and positions. The function $\lambda_k(x_{d,i}^u)$ is the linear regression of the scale parameter: $\lambda_k(x_{d,i}^u) = \exp(\omega_{d,k}^T x_{d,i}^u)$. It can be regarded as the pseudo-mean of the conversion delays.

Second, we calculate the probability that a user does not convert in the observation window ($Y = 0$). As expressed in Equation (4), if the conversion does not occur in the observation window, all ads in the browsing paths must have failed to trigger the conversion. There are two possibilities: the user will never convert, or the user will convert later. By the law of total probability, we can thus write the probability that an ad exposure $\{a_i^u, t_i^u, x_i^u\}$ fails to trigger a conversion before T^u as:

$$\begin{aligned} &\Pr(Y = 0|X_c = x_{c,i}^u, X_d = x_{c,i}^d, E = e_i^u) = \\ &\Pr(Y = 0|C = 0, X_d = x_{d,i}^u, E = e_i^u) \Pr(C = 0|X_c = x_{c,i}^u) + \\ &\Pr(Y = 0|C = 1, X_d = x_{d,i}^u, E = e_i^u) \Pr(C = 1|X_c = x_{c,i}^u) . \end{aligned} \tag{8}$$

It is obvious that

$$\Pr(Y = 0|C = 0, X_d = x_{d,i}^u, E = e_i^u) = 1 , \tag{9}$$

and

$$\Pr(C = 0|X_c = x_{c,i}^u) = 1 - p(x_{c,i}^u) . \tag{10}$$

Furthermore, the probability of the conversion delay being longer than T^u is:

$$\begin{aligned}&\Pr(Y=0|C=1,X_d=x^u_{d,i},E=e^u_i)\\&=\int_{e^u_i}^{\infty}\varphi_k(x^u_{d,i},t)dt=S_k(x^u_{d,i},t)\end{aligned}\tag{11}$$

which is exactly the survival probability.

Combining Equations (6), (9), (10) and (11), the probability of not observing a conversion can be written as:

$$\begin{aligned}&\Pr(Y=0|X_c=x^u_c,X_d=x^d_c,E=e^u_i)\\&=1-p(x^u_{c,i})+p(x^u_{c,i})\int_{e^u_i}^{\infty}\varphi_k(x^u_{d,i},t)dt\\&=1-p(x^u_{c,i})\left[1-S_k(x^u_{d,i},e_i)\right]\ .\end{aligned}\tag{12}$$

Finally, the log likelihood of the observed data is:

$$L(\phi)=\sum_{u:y^u=1}^{U}\log\Pr(Y=1|b^u)+\sum_{u:y^u=0}^{U}\log\Pr(Y=0|b^u).\tag{13}$$

3.3 Parameter Estimation

The above proposed model has three parameters: α the shape parameter, ω_c for the intrinsic conversion rate, and ω_d for conversion delay. It is possible to use the gradient optimization algorithms to train the model. However, the conversion delays would only be observed in positive users (those who converted), and the complicated form of the conversion rate $\Pr(Y=1|b^u)$ (Equation (3)) makes it difficult to calculate the gradient of the log likelihood function (Equation (13)). Therefore, we untangle the conversion rate $p(x^u_{c,i})$ and the conversion delay $\varphi_k(x^u_{d,i},d^u_i)$ in the likelihood function, and then estimate them in two stages.

First, we estimate the parameters for conversion delay. The shape parameter α_k and the scale weight $\omega_{d,k}$ which are only associated with $\varphi_k(x^u_{d,i},d^u_i)$ would be optimized with the positive users for simplification. The log likelihood of the observed conversion delays is:

$$\begin{aligned}L(\phi_d)=&\sum_{u:y^u=1}^{U}\sum_{i}^{l_u}\log\Pr(D=d^u_{d,i}|X_d=x^u_{d,i},C=1)\\&\Pr(C=1|X_c=x^u_{c,i})\\=&\sum_{u:y^u=1}^{U}\sum_{k}^{K}\sum_{i:a^i=k}^{l_u}\log\varphi_k(x^u_{d,i},d^u_i)\ ,\end{aligned}\tag{14}$$

where $\Pr(C=1|X_c=x_c)=1$ for all the converted users. In practice, we carry out the parameter estimation using maximum a posteriori (MAP) with a *Gamma* prior for λ_k and a *Gaussian* prior for ω_k.

Second, we estimate the intrinsic conversion rate $p(x^u_{c,i})$ by optimizing the log likelihood (Equation (13)) with respect to the parameter of ω_c. In this stage, the other two parameters are fixed at the values obtained from the previous stage.

The objective functions for both stages are twice differentiable and unconstrained, so any gradient optimization algorithm could be employed. In our experiments, we have used L-BFGS, one of the most popular optimization algorithms for parameter estimation in machine learning.

3.4 Multi-Touch Attribution

Once the model parameters have been estimated, we can then use the hazard rate $h(b^u_i)$ to measure the the influence of an ad exposure $\{a^u_i,t^u_i,x^u_i\}$ on the conversion when the conversion delay is $d^u_i=T^u_c-t^u_i$. The hazard rate can be calculated as:

$$h(b^u_i)=\frac{\alpha_k}{\lambda_k(x^u_{d,i})}\left(\frac{d^u_i}{\lambda_k(x^u_{d,i})}\right)^{\alpha_k-1}\ ,\tag{15}$$

where the timestamp of the conversion is T^u_c and $d^u_i=T^u_c-t^u_i$ is the conversion delay of the ad exposure b^u_i.

Next, we calculate the contributions of an ad exposure $\{a^u_i,t^u_i,x^u_i\}$ and that of a channel k to the conversion as:

$$att^u_i=\frac{h(d^u_i)}{\sum_{i=j}^{l_u}h(d^u_j)}\quad\text{and}\quad att^u_k=\sum_{a^u_i=k}att^u_i\ .$$

The amount of contribution would be affected by the influence strength, its decay speed and the exposure time.

3.5 Conversion Prediction

The attribution models learned from the observed data can be applied to conversion prediction which is of great significance for advertisers to assess the potential benefit of advertising campaigns and revise the allocation of their budget to different advertising channels. Moreover, conversion prediction provides a feasible indirect way to evaluate attribution models: a more accurate attribution model is likely to yield more accurate conversion predictions. Thus we are able to evaluate attribution models objectively in spite of the lack of ground truth for conversion attribution.

Suppose $\{a^u_i,t^u_i,x^u_i\}_{i=1}^{l_u}$ is the ad browsing path of the user u, and we would like to predict the conversion rate in the time window T'. To this end, we need to consider two problems: whether the user will convert, and when she will convert.

Whether the user will convert is determined by the intrinsic conversion rate of the user:

$$\Pr(C=1|X_c=x^u_{c,i})=p(x^u_{c,i})\ ,$$

where $p(x^u_{c,i})$ that is time-independent has already been defined in Equation (6). If the probability $p(x^u_{c,i})$ is very low, the user could be deemed to have no interest in these ads and would never convert. If the user does convert ($C=1$), the conversion time T^u_c is determined by the distribution of conversion delay $\varphi_k(x^u_{d,i},d^u_i)$:

$$\Pr(D=d^u_i|X_d=x^u_{d,i},C=1)=\varphi_k(x^u_{d,i},d^u_i)\ ,$$

where $\varphi_k(x^u_{d,i},d^u_i)$ has already been defined in Equation (7).

If the user has encountered just one ad a^u_i by time t^u_i, the probability that the conversion would occur within the time window T is:

$$\begin{aligned}&\Pr(Y=1|X_c=x^u_c,X_d=x^d_c,E=e^u_i)\\=&p(x^u_{c,i})\int_0^{e^u_i}\varphi_k(x^u_{d,i},t)dt\\=&p(x^u_{c,i})\left[1-S_k(x^u_{d,i},e_i)\right].\end{aligned}\tag{16}$$

Considering the joint effect of all the relevant ads in the ad browsing path, the probability that the conversion would

occur within the time window T' could be formulated as:

$$\begin{aligned}&\Pr(Y=1, T_c < T' | b_u)\\ =&1 - \prod_{i=1}^{l_u} \left[1 - \Pr(Y=1 | X_c = x_c^u, X_d = x_c^d, E = e_i^u)\right]\\ =&1 - \prod_{i=1}^{l_u} \left\{1 - p(x_{c,i}^u)\left[1 - S_k(x_{d,i}^u, e_i)\right]\right\},\end{aligned} \tag{17}$$

which is similar to the conversion rate given by Equation (3).

4. EXPERIMENTS

To evaluate the proposed model, PMTA, we have designed and conducted a series of experiments. First, we introduce the experimental setup, including the dataset and the evaluation methods. Second, we interpret the parameters of the proposed model and test the goodness of fit of the Weibull distribution for modelling conversion delay. Third, we measure the performance of the proposed model for the task of conversion prediction and compare it with the existing models. Lastly, we discuss the attribution results generated by the proposed model.

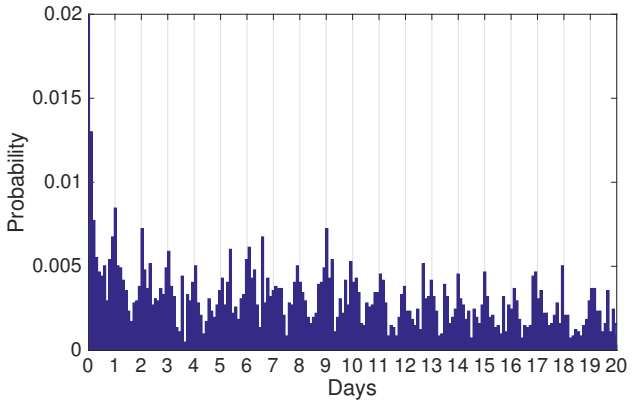
4.1 Dataset

Our experiments have been carried out on a large real-world dataset provided by Miaozhen[4], a leading marketing company in China. It is the log of an advertising campaign that was running from May 1, 2013 to June 30, 2013. There are about 1.24 billion data records in the log, each of which describes an ad exposure (that a user viewed or clicked an ad through a certain advertising channel) by the following attributes: the timestamp, the user ID, the channel ID, the advertising form, the website address, the type of operation system and browser, etc. In addition, the dataset also contains information about each conversion including the timestamp of conversion and the corresponding user ID. Given all such information, we could re-construct the full ad browsing path of each user in this advertising campaign, which consists of the chronological sequence of the exposed ads, user actions (impressions or clicks), channels (the display forms and positions of the ads), and conversions.

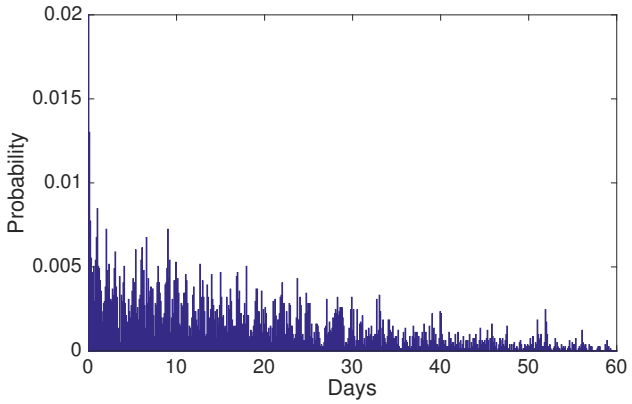
The raw dataset does contain a lot of noise, though. To ensure the reliability of experimental results, we have cleaned it by removing the following data records: (i) the users who have viewed only 1 relevant ad throughout the campaign as those users probably have not been influenced by the campaign; (ii) the re-conversions within 7 days as such short-term re-conversions are probably not triggered by the ads after the previous conversion; (iii) the ad exposures which are not the last 20 ones in the browsing path, as on average only 12 ads are viewed or clicked before a conversion.

In total, there are roughly 59 million users and 1044 conversions available. This advertising campaign involved 2498 channels with 40 forms (e.g. iFocus, Button, Social Ad) and 72 websites (e.g. video websites, search engines, social networks). As only about 0.01% of all users were ever converted, we randomly sampled 10% of negative examples for the training of our model.

Thus, we end up with 14856 observed conversion delays of all positive users. The distribution of the observed conversion delays over 20 days and 60 days are shown in Figure 4a and Figure 4b respectively. The daily cyclic pattern exhibited in the former graph indicates that users are likely to view or click ads in the same peak hours of the day. The long-tailed pattern exhibited the latter graph confirms that it is indeed reasonable to model conversion delays as lifetime durations.



(a) 20 days



(b) 60 days

Figure 4: The distribution of observed conversion delays.

4.2 Baseline Methods

The proposed PMTA model is empirically compared with the following existing attribution models.

- **AdditiveHazard**: the state-of-the-art model. It uses the additive hazard rate to measure the influence of ad exposures on user conversions, though it does not take the intrinsic conversion rate of a user into account [20].
- **Simple Probability**: a straight-forward attribution method. It simply takes the observed conversion probability of each channel and calculates the conversion rate of a user with Equation (3). Given the ad browsing path of user u, the probability of conversion is:

$$\Pr(Y=1 | \{a_i^u\}_{i=1}^{l_u}) = 1 - \prod_i^{l_u} \left(1 - \Pr\left(Y=1 | a_i^u = k\right)\right).$$

- **Time-aware**: a time-aware model for conversion prediction. It is based on last-touch attribution and focused on conversion delay [3].

[4]http://www.miaozhen.com/en/index.html

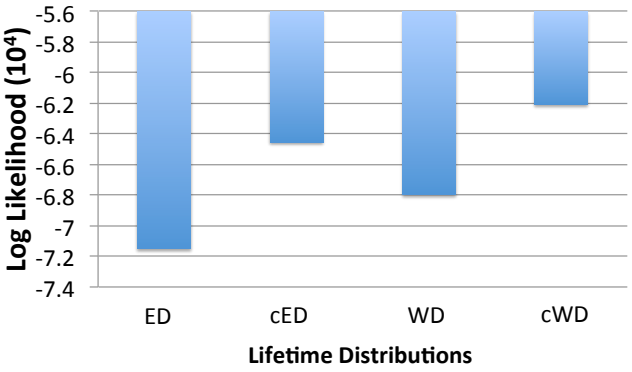


Figure 5: The log likelihood of observed conversion delays under the assumption of four different distributions.

- **Logistic Regression**: the earliest data-driven multi-touch attribution model in the research literature [14].

4.3 Choice of Probability Distribution

In our proposed model, we assume that conversion delay follows the Weibull distribution which is context-aware, i.e., adaptive to the context. To find out whether the context-aware Weibull distribution is really a good choice or not for this purpose, we use the log likelihood of data to check how well the Weibull distribution based model fits the observed conversion delays. For all converted users, the log likelihood of their conversion delays can be calculated as:

$$\begin{aligned} L(\Theta) &= \sum_{u}^{U}\sum_{i}^{l^u}\sum_{k}^{K} \Pr\left(D = d_i^u | X_d = x_{d,i}^u, C = 1\right) \\ &= \sum_{u}^{U}\sum_{i}^{l^u}\sum_{k}^{K} \log \varphi_k(x_{d,i}^u, d_i^u) \ . \end{aligned} \tag{18}$$

Let us compare four distributions: the standard exponential distribution (ED), the context-aware exponential distribution (cED), the standard Weibull distribution (WD) and the context-aware Weibull distribution (cWD). The experimental results are shown in Figure 5. First, we can see that the Weibull distribution always performs better than the exponential distribution (a special case of the Weibull distribution with the shapes parameter $\alpha = 1$). This means that the shapes parameter α plays an important role in modeling conversion delays and the decay speed of a conversion changes with time. Second, the contextual information is indeed helpful in fitting the exponential distribution or the Weibull distribution to the observed data. A plausible explanation is that the context-aware distribution with a self-adapting scale parameter (the pseudo mean value of the conversion delays) is more flexible than the distribution with a fixed scale parameter. These finds are also supported by the oscillating and irregular shape of the distribution curve for conversion delays, as shown in Figure 4.

4.4 Interpretation of Model Parameters

In the proposed PMTA model, we use the context-aware Weibull distribution to describe conversion delays, and then use the hazard rate to measure the influence of an ad exposure to the conversion. Specifically, the shape parameter α_k reflects how fast the influence changes over time, while the scale parameter $\lambda_k(x_{d,i}^u)$ (estimated by a linear regression of the context variable $x_{d,i}^u$) reflects how long most conversion delays are.

Table 1 shows three channels with the highest α_k and another three channels with the lowest α_k in our dataset. The information for each channel includes its ID, type, website, and the value of α_k. It is clear that the shape parameter α_k has values less than 1 for all the channels. As we have explained earlier in Section 3.1, this means that the influence of an ad exposure always fades away over time. The smaller α_k is, the faster the influence decreases. The three channels with the lowest α_k have the fastest speed of decay for the influence of their ads. It turns out that they are all search engines, which suggests that the effect of a paid search ad may disappear very quickly. This is probably because a paid search ad is initiated by a user and the decision whether to purchase will usually be made immediately after the user visits the landing page of the ad.

Furthermore, two example ad exposures are given for each channel in Table 1. For ad exposures, we show the corresponding user ID, the scale parameter $\lambda_k(x_{d,i}^u)$, and also some hazard rates. The value of the scale parameter $\lambda_k(x_{d,i}^u)$ is determined by the context variable $x_{d,i}^u$ and its weight $\omega_{d,k}$. The context variable $x_{d,i}^u$ represents the contextual information of an ad browsing path $\{a_i^u, t_i^u, x_i^u\}_{i=1}^{l_u}$ including the features of each ad exposure (impression or click) a_i^u. Although the weight $\omega_{d,k}$ holds the same value for all ad exposures in a channel, the scale parameter $\lambda_k(x_{d,i}^u)$ may have different values for each individual ad exposure. To demonstrate how the influence of an ad exposure varies with time, we show the value of the hazard rate $h(d)$ for four different conversion delays ($d = 0.001$, $d = 0.1$, $d = 24$, and $d = 240$), where the time d is in the unit of hours. It can be seen that when the shape parameter α_k is large (i.e., the decay speed is slow), the hazard rate is relatively small for short conversion delays and it does not decrease very quickly over time; on the contrary, when the shape parameter α_k is small (i.e., the decay speed is fast), the hazard rate is relatively large for short conversion delays and it decrease very quickly over time.

4.5 Conversion Prediction

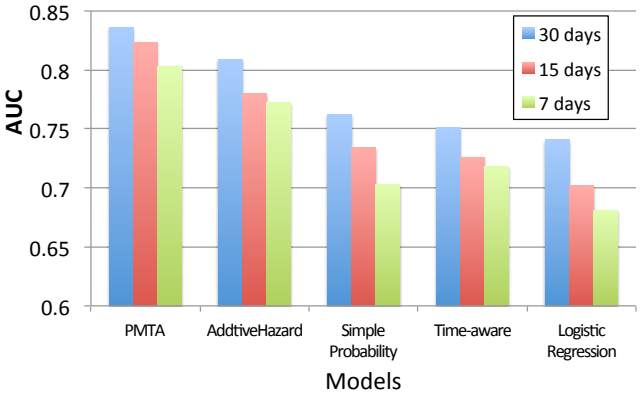
The ability of making conversion predictions based on user behavior data is of great significance for advertisers to assess the performance of their advertising campaign and revise the allocations of their budget to various advertising channels. Furthermore, since there is no ground-truth data available for conversion attribution, we have no direct way to quantitatively measure the effectiveness of an attribution model. A common practice in previous research to get around this obstacle is to use conversion prediction as a feasible indirect way to evaluate and compare different attribution models objectively, as we can reasonably assume that a more accurate attribution model is likely to yield more accurate conversion predictions. .

For a user u, given her ad browsing path $\{a_i^u, t_i^u, x_i^u\}_{i=1}^{l_u}$, we would like to predict her conversion rate (in a specified upcoming period) using our proposed PMTA model as well as the four baseline models described in Section 4.2. In our experiments, the length of the upcoming period is set to be 30, 15 and 7 days. The performance measure is the well-known AUC metric, and the reported results are generated by 4-fold cross-validation (over the users).

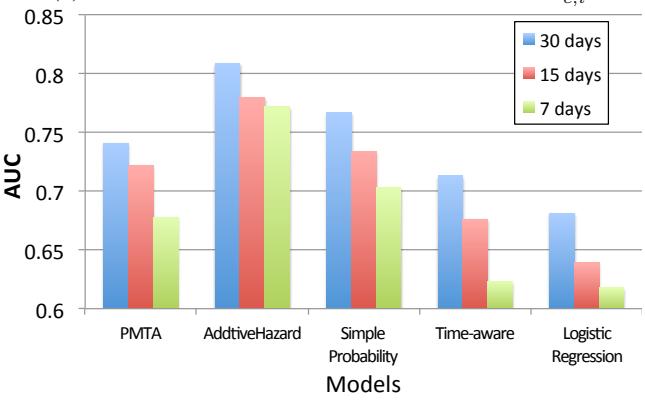
It is notable that search ads are quite different from other

Table 1: The six channels with the highest or the lowest α_k, each of which has two example ad exposures.

Channel	Type	Website	α_k	User ID	$\lambda_k(x^u_{d,i})$	$h(0.01)$	$h(0.1)$	$h(24)$	$h(240)$
100281341	Banner	`Portal 1`	0.832	m117372390	67.39	0.0543	0.0369	0.0147	0.0100
				m154171211	108.42	0.0366	0.0248	0.0099	0.0067
100281056	SEM	`Search 1`	0.789	m346205472	312.62	0.0224	0.0138	0.0043	0.0027
				m536353	421.74	0.0177	0.0109	0.0034	0.0021
100281089	SEM	`Search 1`	0.771	m172805513	13.48	0.2979	0.1758	0.0501	0.0296
				m172805513	4.72	0.6690	0.3949	0.1126	0.0664
100281075	SEM	`Search 1`	0.014	m191937311	9.46	1.2719	0.1314	5.9099e-04	6.1035e-05
				m172805513	89.78	1.2325	0.1273	5.7266e-04	5.9142e-05
100281055	SEM	`Search 1`	0.013	m346220612	0.42	1.2383	0.1276	5.7092e-04	5.8826e-05
				m64923961	0.81	1.2278	0.1265	5.6606e-04	5.8326e-05
100242476	SEM	`Search 1`	0.0032	m23126883	0.35	0.3164	0.0319	1.3515e-04	1.3615e-05
				m23142333	0.72	0.3157	0.0319	1.3484e-04	1.3584e-05



(a) The search ads are excluded in the feature set $x^u_{c,i}$.



(b) The search ads are excluded in the feature set $x^u_{c,i}$.

Figure 6: The experimental results of conversion prediction.

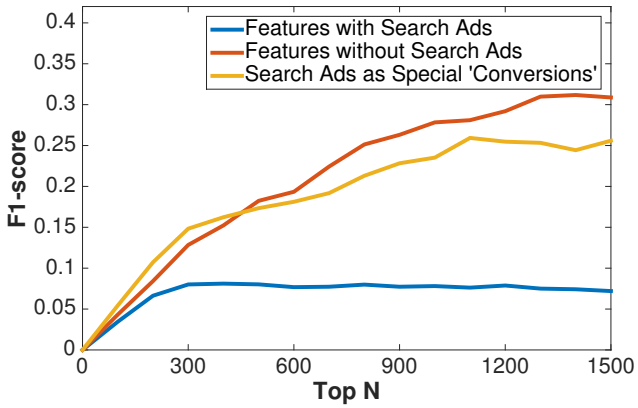


Figure 7: Dealing with search ads in three different manners.

types of ads: as search engines would only return an ad to the user if it is somewhat relevant to the current query submitted by that user, when a search ad is presented to the user she probably is more or less interested in that kind of stuff. Therefore we many not want to deal with search ads in the same way as we handle display ads. The above hypothesis has been examined in the conversion prediction experiments.

First, we use the feature set x^u_i without search ads to predict the conversion rates. As we can see from Figure 6a, the proposed PMTA model outperforms all the other attribution models. The second best model is the AdditiveHazard model, which is also a multi-touch model but does not consider the intrinsic conversion rate of users. The comparison of these two models reveals that eventual conversions are heavily affected by the intrinsic conversion rate. The Logistic Regression model is the worst performing one among the five attribution models put to the test. Furthermore, it is clear that the prediction of conversion rate is more difficult over a short upcoming period than over a long upcoming period. When the period is short, it might be too early to tell whether a conversion will occur or not before the period ends. The performance of the Simple Probability model and the Logistic Regression model decline rapidly when the period becomes shorter. Interestingly, the Simple Probability model can work better than much more complicated models, the Time-aware (last-touch) model and the Logistic Regression model, when the period is not too short.

Then we add search ads to the feature set x^u_i and predict the conversion rates again. The results are shown in Figure 6b. Comparing it with Figure 6a, we are surprised to see that search ads would actually hurt the performance of conversion prediction. One one hand, the AUC scores of the PMTA model, the Time-aware model, and the Logistic Regression model all drop a lot. One the other hand, the AUC scores of the AdditiveHazard model and the Simple Probability model stay unaffected, but it is only because the context features of a user are not used in these two models.

Finally, we conduct an extra experiment to test whether search ads should be deemed to be the consequence of previous ads rather than the cause of the eventual conversion.

Making use of the proposed PMTA model, we consider three possible ways to deal with search ads (including their impressions and clicks) for conversion prediction: (i) including search ads in the feature set x_i^u; (ii) excluding search ads in the feature set x_i^u; and (iii) treating clicked search ads not as features but as special "conversions", i.e., the consequence of previous ads. This time we use the F_1 score to measure the performance. Figure 7 shows the F_1 scores for the top N users who have the highest probabilities to convert within 30 days. If search ads are included in the feature set, the F_1 score does not go beyond 0.07, which suggests that search ads are not really correlated with the conversion rate. More importantly, the best performance is achieved when clicked search ads are treated as special "conversions". Therefore, we can say that the search ads brought to the user by search engines due to their relevance to that user's current query are probably caused by the other ads previously seen by that user. It has also been found that search ads often terminate the path to conversion, which implies that many users visiting the landing pages of search ads would not really go ahead to 'checkout'.

4.6 Attribution Analysis

Now let us analyze the attribution results generated by those five models including our own PMTA. The pre-defined time window (period) is set to 30 days. It is difficult to interpret the attribution to 2498 anonymous channels, so we focus on the attribution of to websites instead. Figure 8a shows the attribution to websites given by each model in comparison, where the websites are sorted by the number of ad exposures. The attributions (credit assignments) made by the proposed PMTA model and the AdditiveHazard model are somewhat similar, and both look plausible. The Simple Probability model tends to assign credits to all websites evenly. The Time-aware model assigns almost all credits to one website `Search 1`, because it is largely a last-touch model. This confirms our speculation that last-touch attribution overestimates the contribution of search ads and ignores the influence of previously viewed display ads. The Logistic Regression model also behaves like a last-touch model.

Since we use the hazard rate to measure the influence of an ad exposure on the conversion, the influence is dependent on time. As we have illustrated in Section 4.4, the shape parameters of all channels are smaller than 1, which implies that the influence of an ad exposure is fading away over time for all channels in our dataset. For search engines, the influence is relatively large for a short conversion interval, but its decay speed is very fast. According to the attribution made by the proposed PMTA model, we should allocate a large amount of credits to the search ads only when the search ads are exposed near to the conversion. If the conversion delay is long, the influence of the search ads might be much more smaller than social ads, vertical ads, or video ads, etc. The reason why the Time-aware model and the Logistic Regression model assign many credits to a search engine `Search 1` could be that they cannot distinguish the the search ads of different conversion delays. The propose PMTA allocates most of the credits to the website `Portal 1`, probably because (i) the decay speed of its influence is very slow, as shown in Table 1; (ii) the amount of its ad exposures is relatively large.

The attribution results of the propose PMTA model have also been compared with those of several typical rule-based models that we have mentioned in Section 1. In Figure 2, we have introduced seven rule-based models, and five of them are compared here. The Last Interaction attribution model is not included, because it is the basis of the Time-aware model which has already been analyzed. The Last AdWords Click attribution model just allocates 100% credits to the website `Search 1`, which is not very interesting. Figure 8b shows the attribution results given by those rule-based models. Comparing it with Figure 8a, we find that the data-driven models and the rule-based models would assign credits to websites quite differently. The Linear attribution model allocates the credits equally to each touch point in the conversion path, so this attribution mechanism could reveal the amount of ad exposures from different websites. It can be seen that using this model `Portal 1` and `Social 1` would receive most of the credits, which implies that these two have most of the ad exposures among all websites.

5. CONCLUSION

The main contribution of this paper is our proposed PMTA model for conversion attribution which takes into account both the intrinsic conversion rate of a user and the conversion delay. It can provide advertisers more granular and actionable insights into the contributions of different advertising channels to the eventual conversions of users.

The PMTA model is fitted to the observed data (conversion rate and conversion delay) rather than relying on simplistic assumptions. Borrowing the techniques from survival analysis, we use the Weibull distribution to describe the observed conversion delay so that the influence of an ad exposure can be measured by the hazard rate of conversion. The PMTA model can be applied to conversion prediction, which is of great significance for advertisers to plan and optimize their advertising campaigns.

The PMTA model has been evaluated on a large real-world dataset. The extensive experiments have proved its superior effectiveness in both conversion prediction and attribution analysis compared with existing attribution models. Moreover, a by-product of our experiments is the surprising finding that on this dataset search ads are not positively correlated with conversions, which suggests that they are probably not the root cause of conversions but just the consequence of previously viewed ads. In future work, we would like to analyse more datasets and investigate whether this property of search ads is a common phenomenon in online advertising.

6. ACKNOWLEDGEMENTS

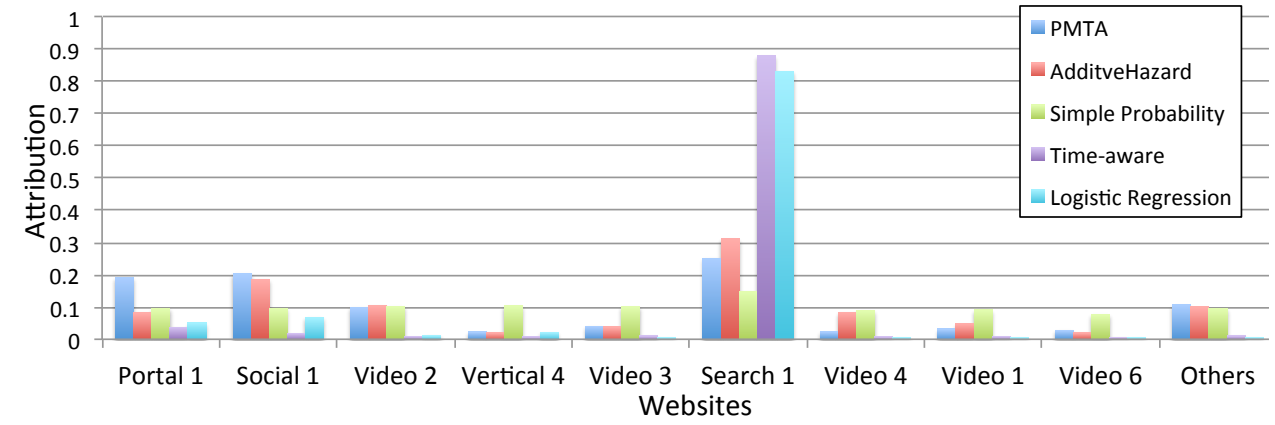
We thank the anonymous reviewers for their constructive comments. This work was partly supported by the NSFC grants (61472141 and 61321064) as well as the Shanghai Knowledge Service Platform Project (ZF1213).

7. REFERENCES

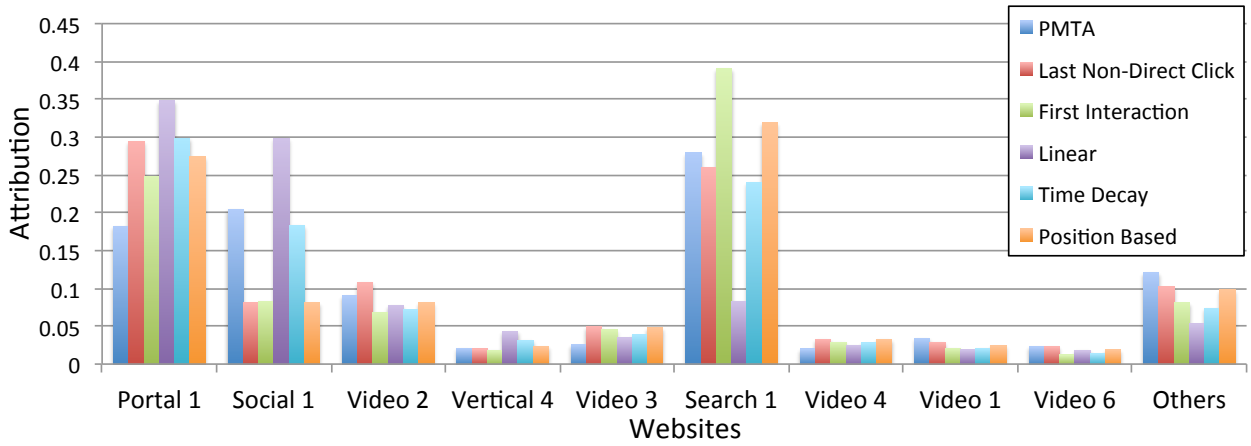
[1] R. N. Bolton. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, 17(1):45–65, 1998.

[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] O. Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International*



(a) Comparing PMTA with data-driven attribution models.



(b) Comparing PMTA with rule-based attribution models.

Figure 8: The experimental results of conversion attribution.

Conference on Knowledge Discovery and Data Mining (KDD), pages 1097–1105. ACM, 2014.

[4] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In *Proceedings of the 6th International Workshop on Data Mining for Online Advertising and Internet Economy*, page 7. ACM, 2012.

[5] F. F. Gönül, B.-D. Kim, and M. Shi. Mailing smarter to catalog customers. *Journal of Interactive Marketing*, 14(2):2–16, 2000.

[6] S. Gupta and V. Zeithaml. Customer metrics and their impact on financial performance. *Marketing Science*, 25(6):718–739, 2006.

[7] D. W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Interscience, 2nd edition, 2008.

[8] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, 3rd edition, 2011.

[9] J. F. Lawless. *Statistical Models and Methods for Lifetime Data*, volume 362. John Wiley & Sons, 2011.

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

[11] P. Manchanda, J.-P. Dubé, K. Y. Goh, and P. K. Chintagunta. The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, 43(1):98–108, 2006.

[12] W. B. Nelson. *Applied Life Data Analysis*, volume 577. John Wiley & Sons, 2005.

[13] S. Richards. A handbook of parametric survival models for actuarial use. *Scandinavian Actuarial Journal*, 2012(4):233–257, 2012.

[14] 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 (KDD)*, pages 258–264. ACM, 2011.

[15] J. Wang and Y. Zhang. Opportunity model for e-commerce recommendation: Right product; right time. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 303–312. ACM, 2013.

[16] D. A. Wooff and J. M. Anderson. Time-weighted multi-touch attribution and channel relevance in the customer journey to online purchase. *Journal of Statistical Theory and Practice*, 9(2):227–249, 2015.

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

[18] D. Zhang, K. Prior, and M. Levene. How long do wikipedia editors keep active? In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym)*, Linz, Austria, Aug 2012.

[19] D. Zhang, K. Prior, M. Levene, R. Mao, and D. van Liere. Leave or stay: The departure dynamics of Wikipedia editors. In *Proceedings of the 8th International Conference on Advanced Data Mining and Applications (ADMA)*, pages 1–14, Nanjing, China, 2012.

[20] Y. Zhang, Y. Wei, and J. Ren. Multi-touch attribution in online advertising with survival theory. In *Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM)*, pages 687–696. IEEE, 2014.