Additional Multi-Touch Attribution for Online Advertising

Wendi Ji, Xiaoling Wang*
Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
3663 North Zhongshan Road, Shanghai, China
wendyg8886@gmail.com, xlwang@sei.ecnu.edu.cn

Abstract

Multi-Touch Attribution studies the effects of various types of online advertisements on purchase conversions. It is a very important problem in computational advertising, as it allows marketers to assign credits for conversions to different advertising channels and optimize advertising campaigns. In this paper, we propose an additional multi-touch attribution model (AMTA) based on two obvious assumptions: (1) the effect of an ad exposure is fading with time and (2) the effects of ad exposures on the browsing path of a user are additive. AMTA borrows the techniques from survival analysis and uses the hazard rate to measure the influence of an ad exposure. In addition, we both take the conversion time and the intrinsic conversion rate of users into consideration to generate the probability of a conversion. Experimental results on a large real-world advertising dataset illustrate that the our proposed method is superior to state-of-the-art techniques in conversion rate prediction and the credit allocation based on AMTA is reasonable.

Introduction

As the growth of computational advertising, targeting techniques make personalized advertising possible. Based on the contextual information and the user feedback data, online advertising systems deliver ads to the users who are most likely to respond. Nowadays companies launch an advertisement campaign through various channels, such as display ad, video ad, social ad, paid search ad and etc. Attribution technology is designed to help marketers understand how particular channels contribute to user conversions, which is now being seen as integral to the future of digital advertising. A promising attribution model is of great help for marketing managers to interpret the influence of channels and optimize their advertising strategies.

In an online advertising campaign, users are exposed to ads with various channels, as illustrated in Figure 1. Suppose that a company X launches an advertising champaign through three channel: display ad, social ad and paid search ad. 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/he searched for products and clicked X's paid ad link at t_3^1. Finally, she/he made a purchase on X's website at time T^1. How shall we evaluate the contribution of the three ads to the conversion?

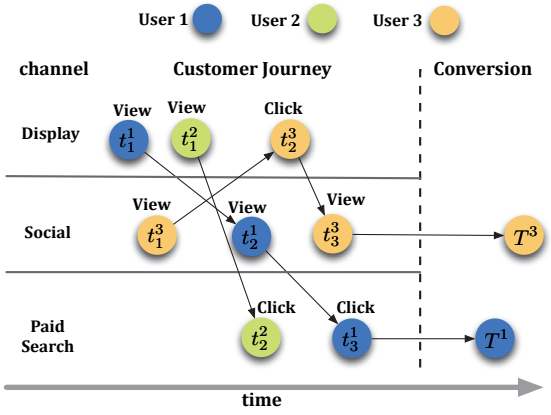

Figure 1: Customer journeys on an advertising champaign. Each journey is composed by a chronological sequence of actions by a user on three advertising channels, including display ad, social ad and paid search ad.

Post-click attribution is one of the earliest and simplest attribution models, which assigns all credit to the last ad clicked before a conversion. It has been considered as the standard attribution model in digital advertising industry. For user 1, if the last click wins, the overall contribution is assigned to the paid search ad and the effects of former viewed ads are totally ignored. Despite its simplicity, this attribution mechanism overestimates the contribution of search ads and neglects the influence of the ads before the last click. In fact, the some queries triggering paid search ads are special conversions due to previously viewed ads. Furthermore, in many cases, user never clicks before conversion. A reliable attribution mechanism should consider the contributions of all relative ads in the consumer journey.

Users' behaviors are caused by the combined effect of the exposed ads within the journey. Multi-touch attribution (MTA) allows marketers to capture real Return of Investment (ROI) for multiple advertising touch points. It has become a significant research topic and has been explored by several online marketing analytics companies (e.g. Google

*The corresponding author is Xiaoling Wang.

Analytics[1], Nielsen[2]). Comparing the ROI based on different MTA models, advertisers evaluate the contribution of different channels and, furthermore, decide how to allocate their budget to various channels in the next stage. Some rule-based MTA models have been proposed in practice, e.g. linear attribution model, time decay attribution model and position based attribution model. However, the main drawbacks of these rule-based models are the subjectivity of hypotheses.

In recent year, several data-driven attribution models have been proposed in computational advertising (Shao and Li 2011; Dalessandro et al. 2012; Zhang, Wei, and Ren 2014) and marketing analytics (Xu, Duan, and Whinston 2014; Wooff and Anderson 2015). However, these existing models only consider either the time-independent conversion rate of a user or the conversion time. First, the influence of an ad is highly related to time, because a user is more likely to be affected by more recent ads. Secondly, the actual conversion also depends on the intrinsic conversion rate of user, because the conversion delay does not exist if the user has no interest in the ads. In fact, the conversions are extremely rare event (the actual conversion rate is as low as about 0.01%), so the conversion prediction solely based on conversion time is biased.

In this paper, we propose a data-driven model for multi-touch attribution and conversion prediction, which is denoted as additional multi-touch attribution model (AMTA). First, we assume that the effect of an ad exposure is fading with time and the effects of multiple ad exposures on the browsing path of a user are additive. Inspired by survival analysis, we use hazard rate to model the effect of an ad exposure upon the conversion, which reflects the effect of an ad exposure to trigger a conversion. The hazard rate of an ad exposure is determined by the influence strength and the decaying speed. It is built for individual ad channel to avoid the bias introduced by different advertising forms and layouts. The distribution of conversion time can be calculated by the additive hazard of all relative ads. Then, we focus on how to predict conversion rate with the proposed AMTA. The conversion prediction based on a MTA model provides great guidance for advertisers to allocate the budget among various channels when starting an advertising champaign. When generating the probability of a conversion, we take both whether the user will convert and when she/he will convert into account. Finally, we evaluated AMTA model using a real-world dataset obtained from Miaozhen[3], a leading marketing technology company in China. The experiments demonstrate the effectiveness of the proposed model in both conversion rate prediction and attribution analysis.

Related Works

In the domain of computational advertising, some recent researches have been devoted to the study of MTA for ad conversions through data-driven approaches. A bagged logistic regression method was proposed to predict the conversion rate based on the viewed ads of a user (Shao and Li 2011), which is the first study in this field. This approach characterize the user journey with the counts of ad exposures and uses the weights to measure the credits of different channels. The drawbacks of this work include: (1) the temporal factor is ignored; (2) the attribution based on logistic regression is difficult to interpret. Dalessandro et al. formulate MTA as a causal estimation problem to achieve interpretable attribution and use the additive marginal lift of each ad to present its credit to conversion (Dalessandro et al. 2012). However, the unbiased estimation of the causal parameters is too complicated to implement and authors therefore developed much simpler approximating methods in practices with subjective assumptions. Zhang et al. proposed an Additivehazard model based on survival theory (Zhang, Wei, and Ren 2014). They modeled the temporal influence of an advertising channel by defining a decay function without the consideration of the intrinsic conversion rate of user and contextual features. However, since it is unknown whether users are interested in the advertising champaign, it is arbitrary to model the impact of an ad exposure. In addition to the extremely sparsity of user conversions, it is even more necessary to consider the intrinsic conversion rate of users. Ji et al. models conversion delay with Weibull distribution and uses the corresponding hazard rate to reflect the influence of an ad exposure (Ji, Wang, and Zhang 2016). This method does not directly measure the combined effect of ad exposure and use one minus the zero effect of all relative ads to generate the multi-touch conversion rate.

There are some researches focusing on MTA in marketing analytics (Gupta and Zeithaml 2006; Li and Kannan 2014). A proportional hazard model was used to predict the conversion time based on the viewed ads of users (Manchanda et al. 2006). It is similar to the logistic regression method (Shao and Li 2011) and the difference is: this one aims at the conversion time, but Shao's model aims at the conversion rate. Inspired by first-touch attribution and last-touch attribution, Wooff et al. used beta distribution to model the influence of an ad exposure, which attributes most credit to the first ad and the last ad (Wooff and Anderson 2015). The common drawback of these models is the ignorance of the intrinsic conversion rate of users. Therefore, these methods fail to provide solely conversion rate prediction.

This work is also related to studies focusing on the time-aware dynamics of ad exposures and conversions based on survival analysis. In marketing analytics, Bolton et al. and Gonul et al. predicted the probability of a customer switching to competitor with proportional hazard models (Bolton 1998; Gönül, Kim, and Shi 2000), where different specifications for the baseline hazard rate are determined by different duration models such as exponential and Weibull. In recommendation system, the same method was used to predict the right time to recommend a product (Wang and Zhang 2013) . Chapelle used exponential distribution to model the delayed feedback of clicked ads (Chapelle 2014). However, these models are all based on last-touch attribution.

The idea of modeling the combined effect of ads by additive hazard is inspired by exciting point process. Yan et al. formulated pipe failure events into a self-exciting stochas-

[1]http://analytics.google.com

[2]http://www.nielsen.com

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

tic process model, which has already deployed as a industrial computational system for pipe failure prediction (Yan et al. 2013). Li and Zha proposed a probabilistic model based on mixtures of Hawkes processes that simultaneously tackles event attribution and network parameter inference to solve the problem of dyadic event attribution (Li and Zha 2013). Yan et al. developed a profile-specific two-dimensional Hawkes processes model to capture the influence from sellers activities on their leads to the win outcome in sales pipeline analytics (Yan et al. 2015). Xu et al. proposed a MTA model based on mutually exciting point process, which considers ad clicks and purchases as independent random events in continuous time (Xu, Duan, and Whinston 2014). Censored data (the event has no occurred) makes survival analysis special and exciting point process only considers the occurrence of event, which is the main difference between them. However, the conversion rate is extremely low for online advertising and it is necessary to take the users who have not converted yet into consideration. The drawback of modeling customer journeys in an advertising champaign with exciting point process is the failure of utilizing unconverted ads. The advantage of our proposed AMTA model is the combination of the survival analysis and exciting point process, which considers both censored data and the additive effects of ads.

Survival analysis

Survival analysis is a widely used approach to have a fine-grained modeling of the observed survival time of products in various fields, including biology, technical reliability, econometrics, sociology, etc (Nelson 2005). As a generic term, the survival time is denoted as the time from the initiating event to the event of interest. We assume that the conversion delay T between an ad exposure and the eventual conversion is the survival time in this work. There are two basic concepts that pervade the whole theory of survival analysis: hazard rate and survival function.

The hazard rate $h(t)$ presents the occurrence rate of the conversion at timestamp t on the condition that the user does not convert before t, which defined (Lawless 2011):

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

The survival function $S(t)$ is defined as the expected proportion of users for which the conversion has not yet occurred by a specified timestamp t. The mathematical connection among the survival function $S(T)$, the hazard rate $h(t)$ and the probability density function $\varphi(t)$ of the survival time t is:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t + \Delta t)}{S(t)} \\ &= -\frac{S(t)'}{S(t)} = \frac{\varphi(t)}{S(t)}. \end{aligned} \tag{2}$$

By integration, using that $S(0) = 1$, we get

$$-\log\{S(t)\} = \int_0^t h(\nu) d\nu, \tag{3}$$

and it follows that

$$S(t) = \exp\left(-\int_0^t h(\nu) d\nu\right). \tag{4}$$

And the probability density function of a conversion occurring at time t is

$$\varphi(t) = h(t) S(t). \tag{5}$$

Therefore, we can define the survival function $S(t)$ and the probability density function $\varphi(t)$ given the hazard rate $h(t)$.

The relationship between Survival function $S(t)$ and the probability density function $\varphi(t)$ is

$$S(t) = 1 - \int_0^t \varphi(\nu) d\nu = 1 - F(t), \tag{6}$$

where $F(t)$ is the cumulative distribution function of $\varphi(t)$.

Additional Multi-touch Attribution

In this paper, we aim to build a probability MTA model to analyze the contribution of each ad exposure to the conversion based on the historical behaviors of users. We assume that the effects of ad exposures on the further conversion are additional and the influence is fading with time. The proposed model is named Additional Multi-touch Attribution Model (AMTA for short).

Additional Effects of Ad Exposures

Before going to the detail of the proposed model, we introduce the notations used in this paper. We denote users as $\{1, ..., U\}$, and the advertising channels as $\{1, ..., K\}$. As shown in Figure 1, we define a behavior $\{a_i^u, t_i^u\}$ as a user u viewing or clicking an ad on an advertising channel a_i^u at some timestamp t_i^u. An ad browsing path b^u of user u is $\{\{a_i^u, t_i^u, x_i^u\}_{i=1}^{l_u}, Y^u, T_c^u\}$, where l_u is the length of the ad browsing path b_u, x_i^u is a set of features, $Y_u \in \{0, 1\}$ indicates whether a conversion has already occurred. If $Y_u = 1$, T_c^u is the conversion time. If $Y_u = 0$, T_c^u is the last timestamp of the observation window. If a user does not convert in an observation window, it is either because the user will never convert or because he/she will convert later. Therefore, an extra variable, $C^u \in \{0, 1\}$, should be considered, which indicates whether a user will eventually convert. Besides, $x_{c,i}^u$, $x_{a,i}^u$and $x_{d,i}^u$ are three 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 be performed when $t_i^u < t < t_{i+1}^u$. $x_{e,i}^u$ determines the effect of the ad exposure $\{a_i^u, t_i^u\}$ and $x_{d,i}^u$ determines its decay speed.

We use the hazard rate to model the additional influence of the ads in the browsing path on the final conversion, which is inspired by the construction of conditional intensity in exciting point process (Aalen, Borgan, and Gjessing 2008). If the user will convert ($C = 1$), the hazard rate of the conversion at time t for user u is:

$$h(t|b^u) = \sum_{t_i^u < t} \alpha_{a_i^u}(x_{e,i}^u) \lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right). \tag{7}$$

In this hazard function, $\alpha_{a_i^u}(x_{e,i}^u)$ denotes the influence strength of the ads from channel k to the conversion, if $a_i^u = k$ and $\lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right)$ denotes its time-decaying kernel.

According to Equation (4), its corresponding survival function is:

$$\begin{aligned} S\left(t|b^u\right) &= \exp\left(-\int_0^t h\left(\nu|b^u\right)d\nu\right) \\ &= \exp\left(-\sum_{t_i^u<t}\alpha_{a_i^u}(x_{e,i}^u)\int_0^{t-t_i^u}\lambda_{a_i^u}\left(\nu, x_{d,i}^u\right)d\nu\right) \\ &= \exp\left(-\sum_{t_i^u<t}\alpha_{a_i^u}(x_{e,i}^u)\Lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right)\right), \end{aligned} \tag{8}$$

where $\Lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right)$ is the integral of the time-decaying kernel $\lambda_{a_i^u}\left(t, x_{d,i}^u\right)$.

Then, given the ad browsing path b^u, the contribution of an ad exposure $\{a_i^u, t_i^u, x_i^u\}$ for the conversion at timestamp t is calculated as:

$$att_i^u = \frac{\alpha_{a_i^u}(x_{e,i}^u)\lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right)}{h\left(t|b^u\right)}. \tag{9}$$

And the contribution of a channel k for the conversion of user u at timestamp t is:

$$att_k^u = \frac{\sum_{t_i^u<t, a_i^u=k}\alpha_{a_i^u}(x_{e,i}^u)\lambda_{a_i^u}\left(t - t_i^u, x_{d,i}^u\right)}{h\left(t|b^u\right)}. \tag{10}$$

Conversion Modeling based on the AMTA

If the user u converts at time t, we can use the proposed AMTA to allocate the contribution of each ad exposure in the path b^u. However, in most cases, the conversion fails to occur within the observation window and the conversion time is censored. In survival analysis, survival function is used to model censored times, e.g. the probability that a patient drops out of the study or a patient still alive at the end of the study. However, the main difference between conversion time analysis and other typical applications of survival time analysis is that: patients will eventually dead, but most of the users will never convert.

Although we aim to model the effect of ad exposures, the influence of the ads to the conversion can not be observed directly. We only observe the users convert within the observation window and the users do not convert within it. If we observe the conversion of a user, we also know when the conversion occurs. If the user does not convert within the time window, he/she may convert later or never convert.

First, we present the probability that a user has converted within the observation window ($Y = 1$). It is obvious that if we observe the observation of user u ($Y^u = 1$), the variable C is also observed: $C = 1$. The probability of user u converting at time t is defined as:

$$\begin{aligned} &\Pr\left(Y = 1, T_c = t|B = b^u\right) \\ =&\Pr\left(C = 1|B = b^u\right)\Pr\left(T_c = t|C = 1, B = b^u\right). \end{aligned} \tag{11}$$

The time independent conversion rate is defined as:

$$\Pr\left(C = 1|B = b^u\right) = p\left(x_{c,i}^u\right), \tag{12}$$

where $t_i^u < t < t_{i+1}^u$. According to Equation (5), the probability of the conversion occurring at time t given the user will convert is:

$$\Pr\left(T_c = t|C = 1, B = b^u\right) = h\left(t|b^u\right)S\left(t|b^u\right). \tag{13}$$

In a similar way, the probability that user u has not converted until T^u can be expressed as:

$$\begin{aligned} &\Pr\left(Y = 0, T_c < t|B = b^u\right) \\ =&1 - \Pr\left(Y = 1, T_c < t|B = b^u\right) \\ =&1 - \Pr\left(C = 1|B = b^u\right)\Pr\left(T_c < t|C = 1, B = b^u\right) \end{aligned} \tag{14}$$

Given the user will convert, the probability of user converting before time t is

$$\Pr\left(T_c < t|C = 1, B = b^u\right) = 1 - S\left(t|b^u\right). \tag{15}$$

Parameter Estimation

In practice, we specify the time-independent conversion rate $p(x_{c,i}^u)$ as logistic function:

$$p(x_{c,i}^u) = \frac{1}{1 + \exp\left(-\omega_c^T x_{c,i}^u\right)}, \tag{16}$$

which is the most widely used in computational advertising industry (Chapelle, Manavoglu, and Rosales 2015).

For each channel, the influence strength $\alpha_{a_i^u}(x_{e,i}^u)$ and the time-decaying kernel $\lambda_{a_i^u}\left(t - t_i^u, x_{i,d}^u\right)$ are specified as:

$$\alpha_k(x_{e,i}^u) = \exp({\omega_{k,e}}^T x_{e,i}^u) \text{ and} \tag{17}$$

$$\lambda_k\left(t - t_i^u, x_{d,i}^u\right) = \gamma_k(x_{d,i}^u)\exp(-\gamma_k\left(x_{d,i}^u\right)\left(t - t_i^u\right)), \tag{18}$$

where $a_i^u = k$ and $\gamma_k(x_k^u) = \exp({\omega_{k,d}}^T x_k^u)$. Then, when $a_i^u = k$, we have

$$\Lambda_k\left(t - t_i^u, x_{d,i}^u\right) = 1 - \exp(-\gamma_k\left(x_{d,i}^u\right)\left(t - t_i^u\right)). \tag{19}$$

Finally, the log likelihood for all users is

$$\begin{aligned} L(\Theta) =& \sum_{u:Y^u=1}^{U}\log\Pr\left(Y = 1, T_c = t|B = b^u\right) \\ &+ \sum_{u:Y^u=0}^{U}\log\Pr\left(Y = 0, T_c < t|B = b^u\right) \\ =& \sum_{u:y^u=1}^{U}\log p\left(x_{c,i}^u\right) + \log h\left(t|b^u\right) + \log S\left(t|b^u\right) \\ &+ \sum_{u:y^u=0}^{U}\log\left(1 - p\left(x_{c,i}^u\right)\left(1 - S\left(t|b^u\right)\right)\right), \end{aligned} \tag{20}$$

where $\Theta = \{\omega_c, \omega_{1,e}, ..., \omega_{K,e}, \omega_{1,d}, ..., \omega_{K,d}\}$. This likelihood is the probability of observing a conversion ($Y = 1$) and its timestamp ($T_c^u = t$) and the probability of not observing a conversion ($Y = 0$) before the time window

($T^u = t$). These probabilities are conditioned on the features x_c, x_e, x_d, the ad exposing time t_i, the conversion time t and the model parameters Θ.

The proposed AMTA has three kinds of parameters: ω_c for the time-independent conversion rate, $\omega_{k,e}$ and $\omega_{k,e}$ for the time-decaying influence of channel k. We implement a gradient descent algorithm as the optimization method for experiments in this paper. The optimization method is taken on the regularized negative log likelihood with respect to parameters of p, λ, and Λ:

$$arg\min_{\Theta} -L(\Theta) + \frac{\mu}{2}\left(\|\omega_c\|^2 + \sum_{k}^{K}\|\omega_{k,e}\|^2 + \sum_{k}^{K}\|\omega_{k,d}\|^2\right), \quad (21)$$

where μ is a regularization parameter.

The objective functions is unconstrained and differentiable, so any gradient optimization algorithm could be employed. In our experiments, we have used mini-batch stochastic gradient descent to reduce the communication cost. The gradients of the negative log likelihood with respect to ω_c, $\omega_{k,e}$ and $\omega_{k,d}$ are:

$$\frac{\partial L(\Theta)}{\partial \omega_c} = \sum_{u:Y^u=1}^{U} \frac{1}{p\left(x_{c,i}^u\right)} \frac{\partial p\left(x_{c,i}^u\right)}{\partial \omega_c} - \sum_{u:y^u=0}^{U} \frac{1 - S\left(t|b^u\right)}{1 - p\left(x_{c,i}^u\right)\left(1 - S\left(t|b^u\right)\right)} \frac{\partial p\left(x_{c,i}^u\right)}{\partial \omega_c},$$

$$\frac{\partial L(\Theta)}{\partial \omega_{k,e}} = \sum_{u:Y^u=1}^{U} \frac{1}{h\left(t|b^u\right)} \sum_{t_i^u<t, a_i^u=k} \lambda_k\left(t - t_i^u, x_{d,i}^u\right) \frac{\partial \alpha_k\left(x_{e,i}^u\right)}{\partial \omega_{k,e}} + \sum_{t_i^u<t, a_i^u=k} \Lambda_k\left(t - t_i^u, x_{d,i}^u\right) \frac{\partial \alpha_k\left(x_{e,i}^u\right)}{\partial \omega_{k,e}} - \sum_{u:y^u=0}^{U} \frac{p\left(x_{c,i}^u\right) S\left(t|b^u\right)}{1 - p\left(x_{c,i}^u\right)\left(1 - S\left(t|b^u\right)\right)} \sum_{t_i^u<t, a_i^u=k} \Lambda_k\left(t - t_i^u, x_{d,i}^u\right) \frac{\partial \alpha_k\left(x_{e,i}^u\right)}{\partial \omega_{k,e}},$$

$$\frac{\partial L(\Theta)}{\partial \omega_{k,d}} = \sum_{u:y^u=0}^{U} \frac{1}{h\left(t|b^u\right)} \sum_{t_i^u<t, a_i^u=k} \alpha_k\left(x_{e,i}^u\right) \frac{\partial \lambda_k\left(t - t_i^u, x_{i,d}^u\right)}{\partial \omega_{k,d}} + \sum_{t_i^u<t, a_i^u=k} \alpha_k\left(x_{e,i}^u\right) \frac{\partial \Lambda_k\left(t - t_i^u, x_{d,i}^u\right)}{\partial \omega_{k,d}} - \sum_{u:y^u=0}^{U} \frac{p\left(x_{c,i}^u\right) S\left(t|b^u\right)}{1 - p\left(x_{c,i}^u\right)\left(1 - S\left(t|b^u\right)\right)} \sum_{t_i^u<t, a_i^u=k} \alpha_k\left(x_{e,i}^u\right) \frac{\partial \Lambda_k\left(t - t_i^u, x_{d,i}^u\right)}{\partial \omega_{k,d}}.$$

Experiments

Dataset

We conduct our experiments on a real-world competition dataset provided by Miaozhen, a leading marketing technique company in China. This dataset includes almost 1.24 billion advertising log of a champaign from May 1st, 2013 to June 30th, 2013. In the dataset, each record describes that a user viewed or clicked an ad through some advertising channel, including the exact time, user ID, channel ID, advertising form, website, the type of operation system and browser, the stability of user ID and etc. In addition, the dataset also provides the conversion data recording the user ID and the exact conversion time. From these data, we can construct the ad browsing journey of a user, including the chronological sequence of the exposed ads, user actions (impressions or clicks), channels (the display forms and positions of the ads) and conversions.

The dataset contains about 59 million users and 1044 conversions. This campaign contains 2498 channels with 40 various advertising forms (e.g. iFocus, Button, Social Ad) and 72 websites (e.g. video website, search engine, social network). The distributions of ad exposures, clicks and the distributions of channel appearances all show long-tailed patterns. Because the advertising log contains lots of noises, in order to get reliable results, we clean the data set by removing some records according to the following rules: 1) Remove the users who views less than 2 ads. We assume that if a user just viewed the ad in a champaign, the ad champaign has no contribution to the conversion of this user. 2) Remove the re-conversions within 7 days because a short-term re-conversion may be not motivated by ads. 3) Remove the ad exposures which are not the last 20 ads in the browsing path. It is because that only 12 ads, on average, are viewed or clicked before a conversion. Since the converted users is about 0.01% of all, we sampled 1% negative users for model training.

Baseline Methods

We compare the proposed AMTA model with the following baselines:

- **PMTA**: the attribution model modeling conversion delay with Weibull distributions and using the corresponding hazard rate to reflect the influence of an ad exposure. This method does not directly measure the combined effect of ad exposure and use one minus the zero effect of all relative ads to generate the multi-touch conversion rate. (Ji, Wang, and Zhang 2016).
- **AdditiveHazard**: the attribution model using additive hazard rate to reflect the influence of relative ads on user conversion. This method does not take the contextual information and the intrinsic conversion rate of users into account (Zhang, Wei, and Ren 2014).
- **Simple Probability**: a straight-forward attribution. We compute the empirical conversion probability of each channel and calculate the probability of conversion of user u as:

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

- **Logistic Regression**: the first data-driven MTA model in computational advertising (Shao and Li 2011).

Table 1: The six channels with the highest or the lowest α_k.

Channel	Type	Website	α_k	γ_k
100234261	Column	Search Engine 1	1.437	0.032
100242089	SEM	Search Engine 1	0.781	0.0078
100242639	SEM	Search Engine 1	0.518	7.89
100262175	Video	Video Site 5	8.1e-6	0.0042
100242450	iFocus	Vertical 7	9.4e-7	0.072
100275296	SEM	Search Engine 1	8.9e-7	0.0038

Table 2: The six channels with the highest or the lowest γ_k.

Channel	Type	Website	α_k	γ_k
100275048	SEM	Search Engine 1	0.027	213.2
100275520	SEM	Search Engine 1	0.003	208.1
100248636	Branded Album	Search Engine 1	0.0042	156.2
100281056	SEM	Search Engine 1	7.8e-5	9.8e-4
100281085	SEM	Search Engine 1	3.2e-5	9.6e-4
100281341	Banner	Portal 1	2.1e-5	7.2e-4

- **Time-aware**: a time-aware conversion rate prediction model based on post-click attribution. It is not an attribution model and focuses on conversion delay (Chapelle 2014).

Interpretation of Model Parameters

In the AMTA model, we use the hazard rate to model the additive influence of the ads on the final conversion. For an ad exposure at time t_i from channel k, the hazard rate at time t is determined by the influence strength $\alpha_k(x^u_{e,i})$ and its time-decaying kernel $\lambda_k\left(t - t^u_i, x^u_{d,i}\right)$. The influence strength and time-decaying kernel of the AMTA model are both associated with the contextual features, as we defined in Equation (17) and Equation (18). The influence strength $\alpha_k(x^u_{e,i})$ is estimated by a linear regression of $x^u_{e,i}$. The time-decaying kernel is determined by both the the decaying speed $\gamma_k(x^u_k)$ and the delay $t - t^u_i$. Here, we leave the features, both $x^u_{e,i}$ and $x^u_{d,i}$, out and focus on how to the influence strength α_k and the decaying speed γ_k determines the effect of channel k.

Table 1 shows three channels with the highest α_k and another three channels with the lowest α_k in our dataset, while Table 2 shows three channels with the highest γ_k and another three channels with the lowest γ_k. The information for each channel includes its ID, type, website, and the value of α_k and the value of γ_k. A big α_k means that the ad exposures from channel k have strong impact on the conversion decision of a user and a big γ_k means that the influence decreases quickly. The three channels with the highest γ_k are all search engines, which suggests that for these channels the effect of a paid search have strong influence on the conversion. On the other hand, the three channels with the highest γ_k are all search engines, which suggests that for these channels 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.

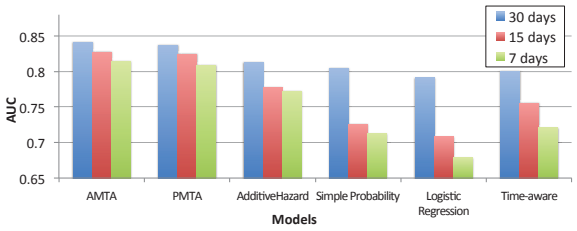

Figure 2: The experimental results of conversion prediction.

Conversion Rate Prediction

The conversion rate is a basic metrics of ad quality in computational advertising. Conversion rate prediction is of great significance for advertisers revise their budget allocation among various advertising channels and deliver the right ad to the right user by ranking the conversion rates. Furthermore, we have no direct way to quantitatively measure the effectiveness of an attribution model, because there is no ground truth in conversion attribution. A common assumption in previous research is that a more accurate attribution model is likely to yield more accurate conversion predictions. Therefore, conversion prediction is always used as an indirect alternative method to evaluate and compare different attribution models.

Here we predict whether a user will convert in a specified upcoming period (30, 15 and 7 days). We use the area under the ROC curve (AUC) to measure the accuracy of conversion rate prediction models, and the results are generated by 4-fold cross-validation over the users. It is notable that search ad is a special advertising type because, in most cases, paid search ads are triggered by queries with clear intentions of users and they should not be treated the same as other channels(Zhang, Wei, and Ren 2014; Ji, Wang, and Zhang 2016). Therefore, we don't use paid search ads in feature $x^u_{c,i}$.

As we can see from Figure 2, the proposed AMTA model performs the best in all examined models. AMTA and PMTA take both the conversion delays and the intrinsic conversion rate of users into consideration, and the AUC value of AMTA is slightly better than PMTA. The next best model is the AdditiveHazard, which does not consider the intrinsic conversion rate of users. Through the comparison of these three models, we can find that the intrinsic conversion rate of users effects their actual conversions. Furthermore, it is obvious that prediction is more difficult over a short period than over a long period, which can be explained that if the elapsed time is too short, it is too early to say a conversion never occur. The performance of the Simple Probability and the Logistic Regression decline significantly when the period is shorten. Interestingly, the Simple Probability model performs better than Time-aware and the Logistic Regression when the prediction period is 30 days.

Attribution Analysis

We next give the attribution analysis of the five different methods. Since the proposed PMA model and AdditiveHazard model consider the time-decaying property, we set the

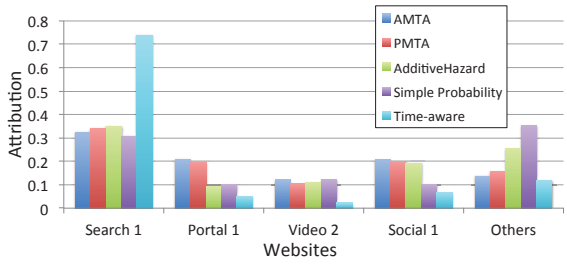

Figure 3: The experimental results of conversion attribution.

pre-defined time window to 30 days. Since it is difficult to interpret the attribution of the anonymous 2498 channels, we here demostrate the attribution of the websites. In addition, Logistic Regression model is not examined in this experiment because the paid search ads is not used the training of Logistic Regression model.

As shown in Figure 3, the proposed AMTA model has the closest comparison with the PMTA model. The difference between these two models is: AMTA assumes the effects of ad exposures are additive and directly models the effects with hazard rate, while PMTA does not directly consider how the effects of ad exposures combine. For the rationality and interpretability of attribution, AMTA is better than PMTA. The only last-touch model, Time-aware model, almost gives all credits to search ads, which is consistent with our assumption that last-touch attribution overestimates the contribution of paid search ads and ignores the influence of other types.

Conclusion

The research proposes an additional multi-touch attribution model for advertising conversions to gain more granular and interpretable insight of the true effects of ad exposure on the conversions. Based on the assumes that the impact of ad exposures is additive and fades with time, we directly use hazard rate to reflect the influence of an ad exposure. In particular, the proposed model considers both the intrinsic conversion rate of a user and the conversion delay. Experimental results show that the proposed model surpasses the existing attribution methods.

Acknowledgements

This work was partly supported by the NSFC grants (61472141 and 61321064) as well as the Shanghai Knowledge Service Platform Project (ZF1213).

References

Aalen, O.; Borgan, O.; and Gjessing, H. 2008. *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Bolton, R. N. 1998. 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.

Chapelle, O.; Manavoglu, E.; and Rosales, R. 2015. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(4):61.

Chapelle, O. 2014. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1097–1105. ACM.

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

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

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

Ji, W.; Wang, X.; and Zhang, D. 2016. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1373–1382. ACM.

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

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

Li, L., and Zha, H. 2013. Dyadic event attribution in social networks with mixtures of hawkes processes. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1667–1672. ACM.

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

Nelson, W. B. 2005. *Applied life data analysis*, volume 577. John Wiley & Sons.

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

Wang, J., and Zhang, Y. 2013. 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*, 303–312. ACM.

Wooff, D. A., and Anderson, J. M. 2015. 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.

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

Yan, J.; Wang, Y.; Zhou, K.; Huang, J.; Tian, C.; Zha, H.; and Dong, W. 2013. Towards effective prioritizing water pipe replacement and rehabilitation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2931–2937. AAAI Press.

Yan, J.; Zhang, C.; Zha, H.; Gong, M.; Sun, C.; Huang, J.; Chu, S.; and Yang, X. 2015. On machine learning towards predictive sales pipeline analytics. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1945–1951. AAAI Press.

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