

Behavior Profiling for Mobile Advertising

Manxing Du
University of Luxembourg
Luxembourg
manxing.du@uni.lu

Radu State
University of Luxembourg
Luxembourg
radu.state@uni.lu

Mats Brorsson
OLAmobile
Luxembourg
mats.brorsson@olamobile.com

Tigran Avanesov
OLAmobile
Luxembourg
tigran.avanesov@olamobile.com

ABSTRACT

Behavioral and targeted profiling of users is an important task in marketing and in the advertising industry. Being able to match a given user profile to an advertising that leads to effective purchases is challenging because of a very tiny proportion of users willing to purchase goods and thus monetize the advertising. With such proportions being less than one percent of the overall user population, efficient feature extraction and modeling techniques are required in order to capture and recognize the potential consumers. This paper proposes a new approach for modeling the observed behavior in a mobile advertising platform, where time related features are correlated with additional system level and campaign related performance statistics. We capture the temporal behavior with Hawkes processes and use the estimated parameters as additional features for predicting if a given user profile will be a revenue generating customer.

Keywords

Display Advertising; Data Mining; Feature Engineering

1. INTRODUCTION

The mobile advertising industry is estimated at 100 billion dollar worldwide for 2016, being driven by the steady increase in tablet and smart phone usage. Targeted profiling of users consists of identifying how users behave in this environment such that relevant advertisement banners can be served. This profiling becomes even more important in the context of Real Time Bidding (RTB) platforms, where online bidding and auctions are performed at millisecond level time scales and accurate predictions are essential for computing the likelihood that a given user profile is a potential purchaser. The bidding strategies design immensely relies on the click through rate (CTR) and conversion rate (CVR) estimation [24]. We describe in this paper our experiences and results in building a profiling engine to improve

the CVR prediction performance for a mobile advertising performance company.

The challenges facing a mobile advertising platform are diverse and range from data mining to efficient data storage and real time system architecture design and implementation. We focus in this paper mostly on the data mining challenges: delivering the most appropriate ads to users for millions of ad clicks. Although many attempts have tried to frame this problem as a supervised learning problem [25, 23], the highly imbalanced classes (less than 0.5% of ad clicks will lead to a purchase) as well as bots/automated scrapers makes this problem inherently challenging. We address this specific problem by modeling user profiles with respect to additional features related to their temporal behavior and global campaign level performance signals. We briefly summarize our contributions below:

- We provide insight into different approaches and their corresponding performances for learning and predicting user behavior using static features like location, software, and time of purchase.
- We integrate global campaign level signals and user profile based signals into the prediction model. The rationale behind it consists in adding exogenous information corresponding to a campaign (and thus summarizing multiple user-profiles) in a similar way to which trading signals are used in trading platform.
- We propose an extended user profile which includes advanced temporal modeling using Hawkes processes. A Hawkes process allows to model non-homogeneous temporal processes, where the intensity at a given time depends on previous observed events with exponentially decaying influences. The rationale behind such a model is that mobile users interact not only with the mobile advertising platform, but also among themselves such that social influence and stimulates more purchases in a group.

Our paper is structured as follows. We start in Section 2 with an introduction to the mobile advertising eco-system and define key concepts for understanding the context and the scope of the problem. Section 3 reviews relevant work on which our contribution builds upon. We provide detailed insights into the features and assumptions in Section 4 and describe our dataset and experiments in Section 5. Metrics used for the assessment of our approach and results are presented in Section 6. Section 7 concludes the paper and highlights the current and ongoing work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BDCAT'16, December 06-09, 2016, Shanghai, China

© 2016 ACM. ISBN 978-1-4503-4617-7/16/12...\$15.00

DOI: <http://dx.doi.org/10.1145/3006299.3006339>

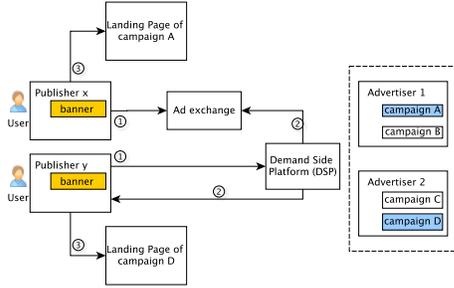


Figure 1: Display advertising system

2. MOBILE ADVERTISING SYSTEM

The eco-system of Internet advertising comprises a complex network of interactions among multiple actors having intertwined business and partnership relations. According to [22], the roles can be reduced to four: user, publisher, ad-exchange, and advertiser. In addition, a DSP can work between ad-exchange and advertisers or between publishers and advertisers as a broker, matching user profiles to ads. It facilitates the cash flows to the advertisers and respectively to the publishers.

Figure 1 demonstrates a simplified display advertising system. The ad slots can be sold in real-time bidding manner or be pre-sold to DSPs. The major difference is in real-time bidding, multiple DSPs will bid to compete for winning an ad impression while in the later case, DSPs only need to send back the optimal campaign in condition to different user profiles and publisher information. When users open an app or a web page, the publisher sends an ad request (step 1) to either ad exchange or DSP. DSP hosts all the campaign information from advertisers, runs prediction model to select the best campaign and sends the redirect URL back to the publisher (step 2). The user will be directed to the landing page if he clicks the ad (step 3). If the user completes any form of purchase later, the advertiser sends back the notification to the DSP as a feedback.

For advertisers, the cost of advertising is measured by different pricing models: Cost per Mille (CPM), the price for showing 1000 impressions; Cost per Click (CPC), the price for each ad click regardless the user purchases anything or not; and Cost per Acquisition (CPA), the price for any form of purchases after the click. Moreover, two metrics are commonly used in the context of user behaviour prediction: the Click Through Rate (CTR), and the Conversion Rate (CVR). CTR measures the ratio of the number of clicks to the number of impressions while CVR represents the percentage of successful purchases over the total number of clicks.

3. RELATED WORK

CTR and CVR prediction has been widely studied in the context of computational advertising. Linear models are widely used for CTR prediction which assume that the features are linearly correlated, ranging from logistic regression (LR) [25], Poisson regression [5], and Bayesian probit regression [8]. However, the individual features may have low correlation with user’s click/convert intention. In [15, 17], authors use feature pairs of pages and ads to learn latent factors through matrix factorization (MF). Furthermore, in [23], the author investigate the advantage of feed forward neural networks in discovering the latent structure in the

high dimensional feature space. Comparing with logistic regression, the deep learning model improves the area under the curve (AUC) for CTR prediction by 0.2%.

However, the sequential nature of purchase and click events are neglected in these studies. The authors of [21] are among the first to model and predict the conversion rate using Hawkes processes. They measure the impacts of the various types of ad clicks prior to the purchases by a mutually exciting point process model. Their model is not tuned to a user profile level granularity, which is the case in our work. A deep introduction to Hawkes processes and their applications can be found in [12, 9]. In [10], the authors model the return time of a user as a survival problem, an instant rate of occurrence is stated as a hazard function, which is similar to our temporal intensity. However, our constraints are much stricter, given the large set of user profiles and huge imbalance between the purchasing users and the non-purchasing ones. Detecting bots and non performing bids is essential for not committing to bid for useless impressions. For this purpose, we use a similar technique to the one described in [7], such that the ad click traffic is profiled based on temporal features.

4. FEATURE AND ASSUMPTIONS

Feature selection is the fundamental step for building a prediction model. We consider three sets of features in this study, defined as static features, temporal features, and feedback features.

4.1 Static Features

The basic feature set can be divided into three groups: user side features, advertiser side features, and publisher side features. User side features include the time stamp of each event, location, operator, device related details such as browser and device type. Campaign ID and the vertical type belong to the advertiser side features, while publisher side attributes are represented by publisher ID and type. A statistical summary about the features can be found in section 5.1.

4.2 Temporal Features

Each user profile’s purchase history can be represented as a stochastic point process $N(t)$, which represents the number of events accumulated until time t . The homogeneous Poisson process [11] is a special class of point process, which assumes the intensity λ of an event is independent of the past and the mean of number of events during a certain time period t can be calculated as λt .

A Hawkes process models the occurrence of an event that depends on previous events. Its conditional intensity function is defined in Equation 1 [12], where $\mu > 0, \alpha > 0, \beta > 0$, t denotes the time since the start of the process, and μ is the baseline intensity, which equals to the intensity of a homogeneous Poisson process. The historical events prior to the current time t are represented as H_t and the time of the i_{th} past event is t_i . In Equation 1, the parameter α measures the intensity increase at time t stimulated by the previous events H_t and β controls how fast the effect decays over time. In other words the further back the event in the process, the less impact it has on future events. Overall, the intensity of a Hawkes process contains the accumulated effects of all the past events prior to the current event, thus it keeps changing over time. In our experiment, the intensity

of Hawkes process is used as the temporal feature.

$$\lambda(t|H_t) = \mu + \alpha \sum_{t_i < t} e^{-\beta(t-t_i)} \quad (1)$$

4.3 Feedback Features

In addition to the temporal features described in the previous section, we also consider the change of conversion rate in the past as an indicator for the future purchase intent. Considering that advertisers notify us the success of conversions as feedback with uncertain delay and also users may not complete the purchase right after clicking the ad, we examine the distribution of the time between two purchases. Let t be the current hour and CVR is computed for hour $(t - 1)$ and hour $(t - 2)$. The CVR change during these two hours for each campaign and each user profile are used as two feedback features, denoting as increase, decrease, and constant.

4.4 Models

In this section, three baseline models are tested for CVR prediction: Logistic Regression (LR), Naive-Bayes (NB), and Random Forest (RF). We use X, y to represent the feature vectors and ground truth labels, respectively. For each feature vector X_i , it contains k different features: $X_i = \langle x_{i1}, x_{i2}, \dots, x_{ik} \rangle$.

LR Logistic regression has been widely used as a linear model for CTR/CVR prediction [19, 13]. It assumes linear dependencies between features: $f = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k$. The probability of having label y_j , given feature vector X_i is $P(y_j|X_i) = \frac{1}{1 + \exp(-f)}$.

NB Different from LR which directly models the conditional probability $P(y|X)$, NB [1] calculates $P(y|X)$ through estimating the joint probability $P(X, y)$ and applying Bayes theorem. NB is based on the assumption that features are independent which adds high bias to the model.

RF is an ensemble classifier [1] which trains multiple decision trees in parallel and averages their probabilistic predictions. Each tree is built by bootstrap sampling a subset of the features and data points. It eliminates the bias of assuming features being either linear dependent or independent as in the previous two models.

5. EXPERIMENTS

5.1 Data Overview

We obtained a real-world dataset from OLAmobile, a global mobile advertising performance company in Luxembourg. The dataset contains one week of ad click and purchase logs of mobile display advertisements in July 2015. Table 1 shows the cardinality of each feature in the dataset. The hour and weekday are extracted from the timestamp of each event. Instead of tracking each individual user’s purchase history, we use the unique combinations of country, browser, operating system, and operator to construct user profiles. By using the aggregated user profile, more purchase events can be collected which facilitate the model training with the temporal features proposed in this study.

During one week, there are ~ 17 M clicks events generated by ~ 17 K user profiles. After each ad click, the notification of conversion may be sent back by the advertisers with

Table 1: Log Data Statistics

Feature	Cardinality
hour	24
weekday	7
country	225
operator	397
device	21018
hardware	4
browser	12
operating system	14
campaign	520
vertical type	5
publisher	1503
publisher type	5

delays which can be up to days. The label of conversion will be added to each click event accordingly. The delay of the conversion has been discussed in a few studies. In [20], the statistics from Yahoo ad exchange shows that 86.7% of the conversions happen within 10 minutes after the ad click, while according to another study with Criteo dataset [4], within one hour of the clicks, only 35% of the conversions can be observed. However, in our dataset the time of each conversion is unknown. We assume the purchase delay and the delay of the purchase notification from advertisers to the mobile advertising platform are constant for each ad campaign. Based on the assumption, we estimate the purchase time to be the same as the click time, which keeps the value of time interval between two purchases to be closer to the reality.

The training set contains 5 days data, and the last 2 days data are used as test set. A well-known public dataset provided by iPinYou [14], one of the biggest DSP in China, is also available. However, in their dataset, 5 out of 9 campaigns do not contain any conversion events and the amount of conversion data is too few to build any model on. Therefore, in the following sections, the results are only based on the dataset from OLAmobile.

The static features in our dataset are categorical features. For example, the feature *country* contains the index of each country, which cannot be treated as numerical features which are ordered by their values. The one-hot-encoding method is used to transform the discrete features to binary vectors for prediction models to process. For instance, if a feature with cardinality of 3: [Firefox, Chrome, Safari], three columns will be needed for the new feature vector. Each categorical variable could be expressed as [0, 0, 1], [0, 1, 0], and [1, 0, 0]. Consequently, after one-hot-encoding, the feature space will tremendously expand to thousands or even millions of dimensions. In our dataset, there are over 600K binary features.

5.2 CVR Change Over Time

The dynamics of CVR throughout the day and a week is shown in this section. Due to the space limit, only the top 5 campaigns ranked by their generated revenues are selected as an example. In Figure 2, the x axis represents the consecutive week days from Monday to Sunday, respectively. It shows that CVR fluctuates over multiple days for campaign number 3287 while the CVRs of other campaigns maintain at a certain value. One explanation is, the campaign 3287 is

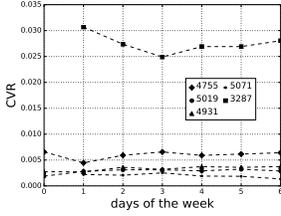


Figure 2: CVR change over a week

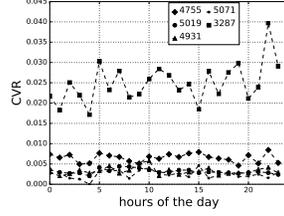


Figure 3: CVR change over a day

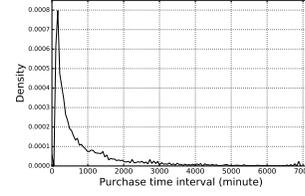


Figure 6: Density distribution of purchase time interval

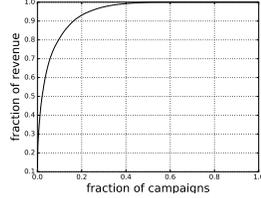


Figure 4: CDF of revenue per campaign

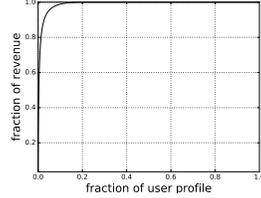


Figure 5: CDF of revenue per user profile

a local campaign which is only available in one country, the other campaigns at the bottom are global campaigns which are launched in over 200 countries, their daily CVR is averaged over all the users. Similar observations are shown in Figure 3, during a day, the CVR of a local campaign varies over each hour while global campaigns have relatively steady conversion rates. It leaves the CVR prediction to be very challenging and requires a more fine-grained solution for each campaign. Given the current time stamp of the ad click, we propose to monitor the trend of CVR change during the past two hours per campaign and per user profile as the feedback features for the prediction model as mentioned in Section 4.3.

5.3 Self Exciting Point Process

In this study, the purchase history of each user profile for each campaign is considered as a series of random point processes. The time of each event is obtained from the time stamp in the log. Fitting each purchase history into both a Hawkes process and a Poisson process helps us to estimate the user’s purchase intent for each campaign. Our hypothesis is the current historical purchase from the same user group increases the intra-group purchase probabilities and the effect decays over time, which is known as the self exciting point process.

Considering that over 500 campaigns are present in our dataset with over 17K user profiles, we first test the Pareto Principle [16], also known as 80-20 rule, to target top campaigns in terms of revenue. Figure 4 shows that the top 10% of campaigns contribute to 80% of revenue. Correspondingly, the top active user profiles are introduced by ranking each user’s total revenue. As is shown in Figure 5, the distribution is highly skewed. The 80% of the revenue is produced by 1.32% of user profiles. Moreover, we found that 7% of the user profiles with most clicks have no purchase records. The behavior of the click-only user profiles is not analysed in this paper, but in the operational setup it can be handled like a bot detection problem. The result suggests it is important and challenging to target the user profiles with the suitable campaign which generates more profit.

Since the point process analysis requires a chain of pur-

chase history, we focus on the campaigns with more purchases and revenues in the following analysis. We selected the data from top campaigns and set a threshold for the number of purchases of each user profile to be at least 100. There are 64 unique combinations of user profiles and campaigns. Figure 6 shows the density distribution of the time between two purchase events for each unique combination of user profile and campaign ID. The first peak in Figure 6 indicates two consecutive purchases from the same user profile arrive within 160 minutes and the probability of having longer intervals decreases. Given the fact that the burst of purchase from the same user group has high probability within 2 to 3 hours, the next step is to model the purchase process to check if it matches a self-exciting process.

The purchase events for each user profile and campaign is evaluated separately. The Akaike information criterion (AIC) is calculated for both Hawkes model (AIC_H) and homogeneous Poisson model (AIC_P), which is used for comparing the fitness of different models [2]. The AIC score is defined as $2K - 2\log(L)$ where K is the number of parameters in the model and L is the maximum likelihood. The model that has lower AIC score fits better. Figure 7 depicts an example of the conditional intensity for a particular user profile purchase history that fits Hawkes process better (with lower AIC score). The x axis shows the minutes since the beginning of the measurement. Each new purchase triggers an increase of more purchases afterwards for short time then decreases back to the background intensity. Another example is shown in Figure 8 where the purchase in the past has negative effect on the subsequent purchases. The intensity drops immediately after a purchase event.

The AIC score is used to compare two models, which cannot show the goodness of fit of each individual model. We further conduct residual analysis [3] for the data which Hawkes process fits better than Poisson process and for the data which Poisson process fits better. For the one-dimensional temporal point process, the residual is defined as: $R(t) = N_t - \int_0^t \lambda(s) ds$, $s < t$, where N_t is the number of events accumulated from time 0 to t , λ is the estimated intensity of the point process model, in our case, it is the Hawkes process, and s is a time point prior to t . Residual is the differences between the real number of events during a certain time and the approximated number of events calculated by the fitted model. Ideally, when the estimated intensity is closer to the real intensity, the residual process should be close to a homogeneous Poisson process with the rate μ estimated by the Hawkes model in Equation 1. Correspondingly, the inter-event time of residual process should be exponential with mean $1/\mu$ [18]. Thus, the better the Hawkes model fits the data, the closer the residual log-plot to be linear. Figure 9 and Figure 10 are the log-plots of residual process for the data in Figure 7 and Figure 8 re-

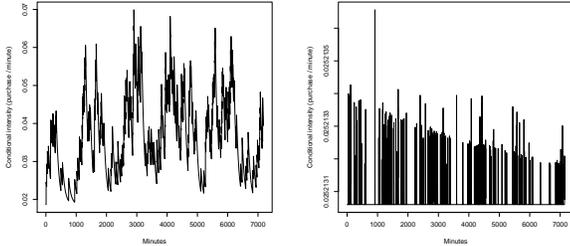


Figure 7: Conditional intensity of a purchase process having $AIC_H < AIC_P$ **Figure 8: Conditional intensity of a purchase process having $AIC_H > AIC_P$**

spectively. It proves that the data in Figure 8 does not fit Hawkes process when the inter-event time gets longer.

Based on the observation, we selected the user profiles with purchase history which can be accurately fitted into a Hawkes process and compute the intensity for every minute since the first purchase event. If the current time is denoted as t minutes since the first purchase, the intensity of $t - 1$ minutes calculated by the fitted Hawkes model is considered as an additional feature for logistic regression. The performance result is summarized in section 6.

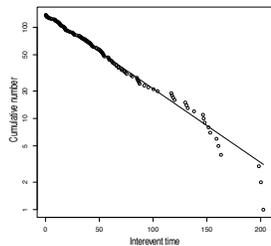


Figure 9: Residual analysis of a purchase process having $AIC_H < AIC_P$

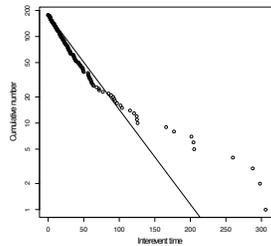


Figure 10: Residual analysis of a purchase process having $AIC_H > AIC_P$

6. EVALUATION

To evaluate the performance of predictive models, the area under Receiver Operating Characteristic (ROC) curve (AUC) is widely used [6], which ranges from 0 to 1. In our dataset, the positive class is purchases and the negative class is the clicks without any following purchase (non-purchases). The dataset is highly imbalanced since the overall CVR is only 0.5%. In this case, accuracy is not an optimal metric to evaluate the performance of the predictor. For example, by predicting all the events as negative class (non-purchases), the predictor reaches 99.6% accuracy. However, its AUC is only 0.5 which is as bad as random guessing. Therefore, AUC has the advantage of insensitivity to imbalanced dataset.

In Table 2, the AUC is computed from training on 5 days of data and testing on the next 2 days as described in section 5.1. The static features in Table 1 serve as base line features. First, the model with the highest AUC score over three basic classification models is selected as the baseline model. The result shows that LR outperforms the other two models, which indicates the features being correlated instead of being independent. NB model with the simple assumption of independent features performs the worst. In the RF

Table 2: AUC Comparison between Baseline Models

Models	AUC
Logistic Regression	0.8154 (0.7109*)
Naive Bayes	0.6362
Random Forest	0.7222*

Table 3: AUC of Feedback Features and Temporal Features

Features	AUC
Static	0.7397
Static + CVR flag	0.7406
Static + CVR flag + Intensity	0.7421

model, each tree selects the square root of the total number of features. Given the large feature space of 600K dimensions in our dataset, we first removed the features with low variance (over 80% of the values are either zero or one) and construct 20 trees for training. For comparison, the same subset of features are used as input for LR model, the corresponding AUC is 0.7109. The AUC with * suggests the model only uses filtered features. Random forest with only 20 trees performs better than LR, however, the training time of RF is 10 times longer than training LR by using a server with 100 GB RAM and 24 cores CPU.

To keep the baseline prediction model to be more efficient with high dimensional features, we select LR as the base line model to compare the performance of adding additional features as proposed in section 4. Since the purchase behaviors of user profiles which can be fitted into Hawkes process are limited, the total clicks generated from these user profiles are about 2 millions over 7 days. These data are chosen to first compute the AUC but using only the static features, followed by adding the feedback feature (CVR change flag) and the temporal feature (intensity) to the LR model. The result depicted in Table 3 demonstrates the importance of considering the temporal nature of the click and purchase events. In practice, we can use logistic regression with temporal features on user profiles which can be fitted to a Hawkes process, while using simple logistic regression without temporal features on the other user profiles.

7. CONCLUSION AND FUTURE WORK

We have addressed in this paper the prediction of user purchases in a mobile advertising context. For this purpose, we propose a new approach that leverages a mix of static and temporal features relating to a user profile and a campaign. Our model groups individual users sharing common features within one user profile and provides predictive solutions for specific campaign related purchases. We have shown how additional and global performance metrics can be used to generate signals that capture the short term trends and identified how these signals can be used for predictive tasks. We have validated our approach on large real world data obtained from a major actor in this industry, covering more than 200 countries and few hundred campaigns. We have evaluated several supervised classification methods and identified their relative strengths and limits for this purpose. We have also investigated the time granularity at which retraining is required in order to capture the inherent dynamics and behavioral shifts occurring in the advertis-

ing markets. The efficient implementation of this approach within a real time bidding platform is our next step. Since our solution requires pairwise modeling of user-profile and campaign level temporal modeling, efficient data structures and computational paradigms are needed for dealing with 20K user profiles and few hundreds campaigns. These challenges require some architectural paradigms that range from distributed message brokers like Apache KAFKA and high speed logistic regression implementations built on GPU systems.

8. ACKNOWLEDGMENTS

Manxing thanks the National Research Fund (FNR) of Luxembourg for the research support under the AFR PPP scheme.

9. REFERENCES

- [1] C. C. Aggarwal. *Data mining: the textbook*. Springer, 2015.
- [2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.
- [3] A. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666, 2005.
- [4] O. Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM, 2014.
- [5] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.
- [6] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [7] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr, and C. Faloutsos. Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278. ACM, 2015.
- [8] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine. In *Proceedings of the 27th International Conference on Machine Learning ICML*, 2010.
- [9] A. G. HAWKES. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.
- [10] K. Kapoor, M. Sun, J. Srivastava, and T. Ye. A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1719–1728. ACM, 2014.
- [11] J. F. C. Kingman. *Poisson processes*. Wiley Online Library, 1993.
- [12] P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes. Papers, arXiv.org, 2015.
- [13] K.-c. Lee, B. Orten, A. Dasdan, and W. Li. Estimating conversion rate in display advertising from past performance data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.
- [14] H. Liao, L. Peng, Z. Liu, and X. Shen. ipinyou global rtb bidding algorithm competition dataset. In *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1–6. ACM, 2014.
- [15] A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 141–149. ACM, 2011.
- [16] M. E. Newman. Power laws, pareto distributions and zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- [17] R. J. Oentaryo, E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 123–132. ACM, 2014.
- [18] R. D. Peng. Multi-dimensional point process models in r. *Department of Statistics, UCLA*, 2002.
- [19] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web, WWW ’07*.
- [20] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012.
- [21] 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.
- [22] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang. Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users. *arXiv preprint arXiv:1206.1754*, 2012.
- [23] W. Zhang, T. Du, and J. Wang. Deep learning over multi-field categorical data. In *European Conference on Information Retrieval*, pages 45–57. Springer, 2016.
- [24] W. Zhang, S. Yuan, and J. Wang. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
- [25] W. Zhang, S. Yuan, and J. Wang. Real-time bidding benchmarking with ipinyou dataset. *CoRR*, abs/1407.7073, 2014.