A LSTM-Hawkes Hybrid Model for Posterior Click Distribution Forecast in the Advertising Network Environment

Sangwon Hwang[1], Inwhee Joe[1*]

1 Department of Computer Science and Engineering, Hanyang University, Seoul, South Korea

* iwjoe@hanyang.ac.kr

Abstract

Popularity forecast for specific user behaviors is a key task in the advertising technology since Key Point of Interests (KPIs) of advertisers are all certain events that users generate. However, as an effect of General Data Protection Regulation (GDPR), it becomes infeasible for Advertising Technology (Ad Tech) individuals to apply machine learning techniques based on user features to the popularity prediction. To overcome this challenge, we propose a new hybrid model for posterior click distribution forecast, named Long Short Term Memory (LSTM)-Hawkes, by combining a stochastic-based generative model and a machine learning-based predictive model. Also, due to innumerable requests and responses for mobile advertisement, easy implementation and computational efficiency are the most critical factors in the Ad Tech. To meet these requirements, we define gradient exponential kernel with just three hyper parameters and minimize residual. The proposed model has been tested with production data. The experimental results show that LSTM-Hawkes reduces the Mean Squared Error (MSE) by at least 27% and up to 79% in comparison to the existing Hawkes Process based algorithm HIP (Hawkes Intensity Process) as well as it improves the forecast accuracy 21.2% in average.

Introduction

For the first time in 2017, global digital advertisement expenditure was 41% of the global advertising market, which exceeded the TV advertisement expenditure, by 6%. Especially the mobile advertisement in digital marketing recorded 37.6% of the global growth rate, which is an impressive increase, enough to draw attention in the worldwide advertisement market [1]. These reports [2–4] substantiate that the mobile advertising market is leading the growth in the global advertising market.

As the mobile advertisement market has rapidly increased, mobile Advertising Technology (Ad Tech) has also become a highly significant area in the advertising market. In 2018, in the United States, Google and Facebook have respectively occupied 38.2% and 21.8% of the mobile advertisement market which is a significant than years before.

However, despite the rapid market growth, there are also some difficulties that Ad Tech faces. Since General Data Protection Regulation [5] has been actively enforced by the EU, machine learning techniques based on consumer information, which could be transferred between Ad Tech individuals before, is now currently restricted in

European countries. Therefore to overcome this barrier, we suggest a generative model based on the Hawkes Process [6], which forms a posterior predictive distribution of a specific event which has occurred in an advertising network. In contrast with the existing machine learning techniques used in Ad Tech [7, 8], we focus on a advertising network's (or an Ad Tech individual) click distribution; not a user's Click Through Ratio (CTR), nor an advertisement's click distribution. This is highly valuable from the advertiser's perspective because a partnership with the right advertising network directly affects the success of their marketing strategy and vice versa (publisher and advertising network).

An advertising network's environment consists of three main parts. First there is the advertiser who wants to advertise, as well as provide the actual advertisement contents. Next, there is the publisher who provides the landing pages for the advertisement contents. Lastly there is the advertising network that connects the advertiser and publisher together. Advertisers are grouped and managed in a system called the 'Demand Side Platform' and publishers are grouped together in a a system called the 'Supply Side Platform'. Fig 1 simply describes the Ad Tech environment and its layers. ***Between layers***, due to the GDPR, user information including advertising IDs or any other unique device IDs from a mobile OS cannot be transferred.

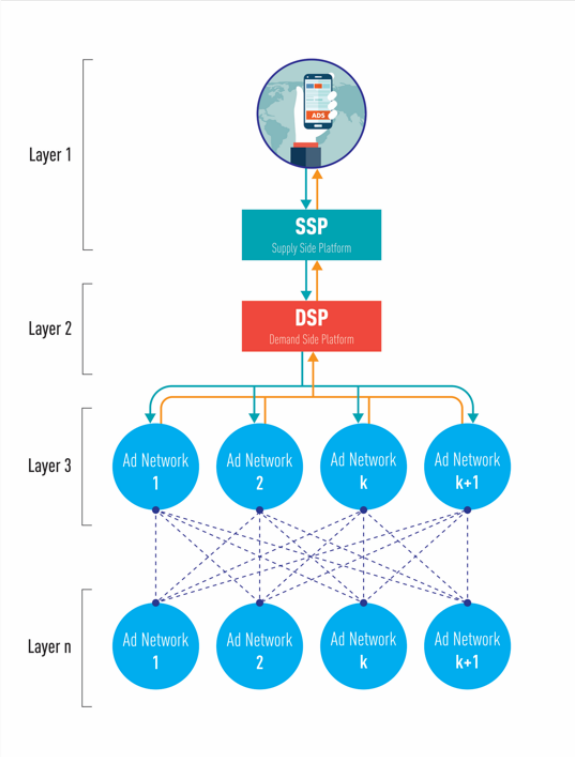


Fig 1. Advertising Network Architecture Using Charles, a debugging proxy server application [9], we captured and analyzed packets from an advertising network and found out that the advertising network environment has a multilayer structure with the cycle flow as shown above. Advertising network environment is composed of advertisers (DSP), suppliers (SSP), and advertising networks (Ad Network).

Materials and methods

Formulation and Preliminary Theory

When a random variable following a Bernoulli process which is the time of occurrence (or success) in Bernoulli trials approaches to Positive infinity, it is infeasible to calculate the probability of a specific event occurrence due to the computational complexity. Therefore, to solve this problem, in modeling a posterior distribution of either prediction or simulation, most conventional statistical approaches have adopted 'Binomial Approximation to Poisson' which helps find probability of an independent trial in various fields [28–30]. However, the 'Binomial Approximation to Poisson' has Memoryless Property [10] because each trial is independent (Independent Increment) and the probability does not change (Stationary Increment). Thus, only event process with non-overlapping intervals can used in the Binomial Approximation to Poisson in modeling a distribution.

In a click event distribution, where the event occurence time is t_i of set $\mathrm{T}_j = \{t_1, t_2, \cdots, t_n\}$, inter-arrival time l_i of set $L_i = \{l_1, l_2, \cdots, l_n\}$ can be written as follows

$$t_{n-1} \leq l_n < t_n \quad \Rightarrow l_n = [t_{n-1}, t_n).$$

And, we can easily find the concurrent occurrences in any advertising click distribution. To solve memoryless property and handle overlapping intervals, we propose a generative model based on a self-exciting process, a type of non-homogenous process, called Hawkes. In this chapter, we provide mathematical induction from binomial distribution to Hawkes process to derive memory kernel of Hawkes.

Binomial Approximation to Poisson. Suppose that a random variable X follows binomial distribution $B(n, p)$, and that the expected value of X is λ. When n is close to infinity, λ approximates to np and by binomial approximation to the Poisson distribution, the probability of X approximates to the probability mass function of a Poisson distribution.

$$f(x) = (n!/x!(n-x)!)(\lambda/n)^x((1-\lambda)/n)^{n-x},$$
$$\lim_{n\to\infty} f(x) = (\exp^{\lambda}\lambda^x)/x! \quad where \;\; X \sim B(n, p).$$

Process. When time unit expands or reduces to by t ($t > 0$), the average event occurrences becomes λt during the updated time unit where the average occurrences per default time unit is $\lambda = np$. Distribution with expanded (or reduced) time unit by t is called Process. And, correspondingly, PDF of the process becomes

$$f(x, \lambda t) = (\exp^{-\lambda t}(\lambda t)^x/x!. \tag{1}$$

Poisson Distribution and Exponential Distribution. Suppose that time period for a specific event to take place is a probability variable t and that T is the time when a specific event X_T takes place. Under this circumstance, the probability of that event X_T occurs after t is equal to that the event X_T does not take place within t. The PDF of Poisson distribution becomes following equation

$$P(t < T) = f(X_T = 0, \lambda t) = (\exp^{\lambda t}(\lambda t)^0)/0! = \exp^{-\lambda t}. \tag{2}$$

Set Eq (2) as $S(t)$ then, the probability of that event X_T occurs is $1 - S(t)$ which is cumulative distribution function (CDF) for the probability variable t. Set this function as $F(t)$ as following Eq (3)

$$F(t) = P(0 \leq T \leq t) = 1 - \exp^{-\lambda t}. \tag{3}$$

Since derivative of CDF is PDF, PDF for the random variable t is

$$f(t) = \frac{d}{dt}F(t) = \lambda \exp^{-\lambda t}. \tag{4}$$

Finally, it is concluded that the probability variable t follows exponential distribution with $f(t)$ and $F(t)$, PDF and CDF respectively.

Non-Homogeneous Poisson Process. In Poisson Process, lambda intensity function $\lambda(t)$ determines the average event occurrence λ per time unit. If $\lambda(t)$ is constant then, the process is homogeneous and if not, the process is non-homogeneous.

In non-homogenous Poisson process, lambda intensity function $\lambda(t)$ is the instantaneous rate at which events occur, defined as the expected rate of arrivals conditioned on associated history up to time t, H_t, divided by during Δt as Eq (3)

$$\lambda(t) = \lim_{\Delta t \to 0} E(N(t + \Delta t) - N(t)|H_t)/\Delta t \tag{5}$$

where $N(T)$, a counting process, is number of event occurrences during T.

Hawkes Process and its Memory Kernel. Since arrival of an event in self-exciting process causes the conditional intensity function Eq (5) to increase, it needs to re-define using memory kernel. The integrated function ϕ is Memory Kernel in the followings where input data are discrete and continuous respectively

$$\lambda(t) = \mu + \sum^{t} \phi(t - T) dN(T),$$

$$\lambda(t) = \mu + \int^{t} \phi(t - T) dN(T).$$

Commonly used memory kernels with Hawkes process are Exponential kernel, original model proposed by Hawkes, and Power Law kernel, frequently used in social network model [11–13], proposed by Ozaki [14].

$$\phi(t - T) = \alpha e^{-\delta(t-T)}, \tag{6}$$

$$\phi(t - T) = k m^{\beta} (t - T + c)^{-(1+\theta)}. \tag{7}$$

LSTM-Hawkes Hybrid Model

The maximum likelihood estimation using exponential Kernel has a problem with residuals which become multiplied by real numbers depending on the batch size. We solved this problem by scaling cumulative intensity value of the kernel with differential coefficient γ that minimizes the residuals, defined Gradient Exponential Kernel Eq (10).

Also, to forecast posterior event time t_i, we adapted Long-Short Term Memory (LSTM) [15] instead of Thinning algorithm [16], dominantly used for sampling with Hawkes, for accuracy enhancement.

In this chapter, we first list all variables and parameters used in the proposed model and define suggested memory kernel. And, in Sampling Method, we show the test result that proves LSTM is more accurate than Thinning algorithm. Lastly, we suggest LSTM-Hawkes Forecast algorithm [Flow Diagram of LSTM-Hawkes] which combines the generative model with LSTM sampling to draw the posterior process of advertising click event. Gradient exponential kernel is defined as Eq (10), and the code implemented with consecutive loops of while, notated in a pseudo code of

Algorithm 1, 2, 3, and 4 where the domain for the memory kernel satisfies $t_i \in eP$ as following

$$\phi(t - t_i) = \gamma\alpha e^{-\delta(t-t_i)} \quad where\ t \in eP. \tag{8}$$

Table 1. Definitions

Parameter	Interpretation
eP_j	Evaluation Point, as argument of $\lambda(t)$
eP	Set of eP_j, equal to domain of t $\lambda(t)$
J	The index of the last element of the set eP
t_i	Observed event time
T	Set of t_i
I	The index of the last element of the set T
$eP_j^{'}$	Evaluation Point, as argument of $\lambda(t)$ for the predictive period
$eP^{'}$	Set of eP_j, equal to domain of t $\lambda(t)$ for the predictive period
$J^{'}$	The index of the last element of the set $eP^{'}$
$t_i^{'}$	Forecasted event time
$T^{'}$	Set of $t_i^{'}$
$I^{'}$	The index of the last element of the set $T^{'}$
τ_n	Distance $[t_i, eP_j)$ such that $t_i \leq eP_j$
γ	Differential coefficient earned by gradient descent using MSE as objective function
μ	Expected value of lambda intensity during a time unit period
θ	S_t of parameters
$L(\theta)$	Likelihood with parameters θ
$l(\theta)$	Log likelihood function with parameters θ
CIF	Cumulative intensity function

Hyper Parameter Estimation. Definition of the lambda intensity function $\lambda(t)$ in non-homogeneous processes was presented in the subchapter 'Formulation and Preliminary Theory' as Eq (5) as well as we derive our own memory kernel as Eq (8). By minimizing negative log likelihood (NLL) over the observed data, hyper parameters for the memory kernel can be achieved. Driven by Rasmussen [17], it is discovered that Eq (9) makes sense by Eq (2), Eq (3) and Eq (4).

$$\begin{aligned} \lambda(t) &= f(t)^{'}/S(t)^{'} = f(t)^{'}/(1 - F(t)^{'}), \\ f(t)^{'} &= \lambda(t)(1 - F(t)^{'}), \end{aligned} \tag{9}$$

where $f(t)$ is PDF of homogeneous Poisson process Eq (4) and $f^{'}(t)$ is conditional probability on associated history up to t_{i-1} which is $f^{'}(t) = \prod_{i=1}^{T} f(t_i|H_{i-1})$. Thus, Eq (2) and Eq (3) can also be re-defined by $f^{'}(t)$ as $S^{'}(t)$ and $F^{'}(t)$ respectively.

Correspondingly, likelihood function for Hawkes can be derived as Eq (10) and its log likelihood function as Eq (11), driven by Rubin [18]

$$L(t_i|\theta) = \prod_{i=1}^{T} f^{'}(t_i|\theta) = \prod_{i=1}^{T} \lambda(t_i|\theta)(1 - F^{'}(t_i|\theta)), \tag{10}$$

$$l(t_i|\theta) = -\int_0^T \lambda(t_i|\theta)dt_i + \int_0^T log\lambda(t_i|\theta)dt_i. \tag{11}$$

Therefore, when eP_j is the domain for the random variable t with PDF $f^{'}(t)$ and the memory kernel is Eq (8), we can derive NLL function for Gradient Exponential Kernel as follows

$$-l(\tau_1, \ldots, \tau_n|\theta) = \mu \cdot \tau_n - \sum_{m=1}^{n} (\gamma\alpha/\delta)(e^{-\delta \cdot \tau_m} - 1) - \sum_{m=1}^{n} log(\mu + \gamma\alpha A(m)), \qquad (12)$$

where $A(m) = \sum_{t_i < eP_m} e^{-\delta(ep_m - t_i)}$ for $i \geq 2$, t_i denotes the event time and A(1) is equal to 0. In our proposed model, optimization is proceeded by minimizing the negative log likelihood Eq (12).

Sampling Method. The evaluation of the sampling result is analyzed by residual distribution. Fig 2 demonstrates that LSTM's sampling residual is conspicuously closer to $f(x) = 0$ than that of Thinning. It is concluded that the prediction of the event time in posterior distribution of LSTM is more accurate than that of Thinning and thinning is preferred to use for simulation based on lambda intensity function.

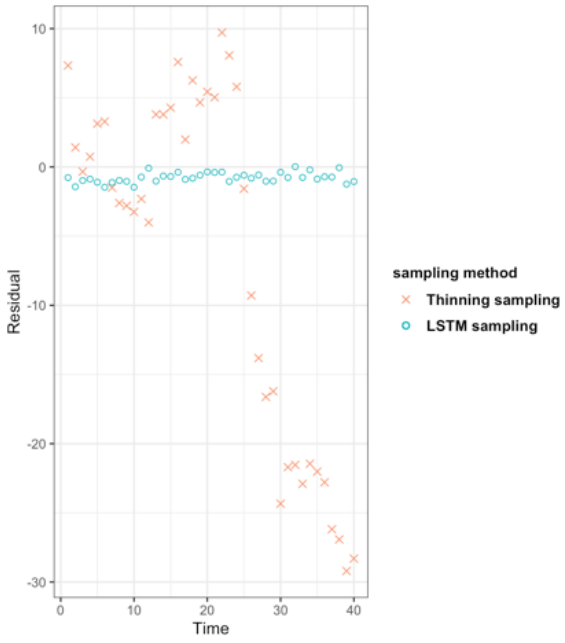
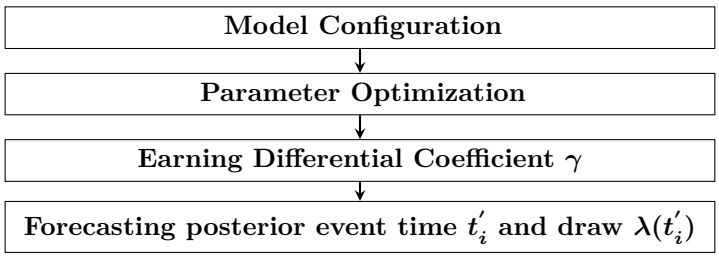


Fig 2. Dispersion of Residuals Residual variance of Thinning is 160.526022455 while that of LSTM is 0.134012729718.

LSTM-Hawkes Model.



Flow Diagram 1. LSTM-Hawkes consist of four simple sub-algorithms.

Algorithm 1: Model Configuration

Data: Observed data set t_i
Result: $\lambda(eP_j)$
$d \leftarrow (t_n - t1)/K$
▷ *Set an equal interval. K is size of set eP*
$eP_1 \leftarrow t_1$
while $(2 \leq j \leq K)$ **do**
  $eP_j \leftarrow t_1 + (j-1)d$
  $j = +1$
$\alpha, \delta \leftarrow \alpha_1, \delta_1$
▷ *Hyper parameter setup*

Algorithm 2: Model Optimization

Data: $\lambda(eP_j)$
Result: α, δ, γ
$l(\theta) \leftarrow Eq(13)$
▷ *Set log likelihood.* $l(\theta)$
while (*Not converged*) **do**
  Run optimization function minimizing $-l(\theta)$
Return Parameters of gradient exponential kernel $\alpha,\ \delta$

Algorithm 3: Earning Differential Coefficient γ

Data: Observed data t_i
Result: γ
for $j = 1;\ j \leq J;\ j++$ **do**
  for $i = 1;\ t_i \leq eP_j;\ i++$ **do**
    $CIF.Append(\lambda(ep_j - t_i))$
$CIF \leftarrow Sum\ of\ CIF\ values\ by\ minute$
$OriginalIntensity \leftarrow Sum\ of\ observed\ event\ number\ by\ minute$
$n \leftarrow size\ of\ CIF$
while $(1 \leq j \leq n)$ **do**
  Residuals $= Original_Intensity_i - CIF_i$
  Optimization of Gradient Descent using MSE as loss function
  Return γ

Algorithm 4: Forecasting posterior event time $t_i^{'},\ \lambda(t_i^{'})$

Data: *CIF of gradient hawkes*
Result: *CIF for* $[t_1,\ t_n^{'})$
$T^{'} \leftarrow\ LSTMSampling(T,\ predictionPeriod)$
$d \leftarrow (t_n^{'} - t_1^{'})/K^{'}$
▷ *Set an equal interval.* $K^{'}$ *is size of set* $eP^{'}$
$eP_1^{'} \leftarrow t_1^{'}$
while $(2 \leq j \leq K^{'})$ **do**
  $eP_j^{'} \leftarrow t_1^{'} + (j-1)d$
  $j = +1$
for $j = 1;\ j \leq J^{'};\ j++$ **do**
  for $i = 1;\ t_i^{'} \leq eP_j^{'};\ i++$ **do**
    $CIF.Append(\lambda(ep_j^{'} - t_i^{'}))$
Return CIF
▷ *CIF for* $[t_1,\ t_n^{'})$

LSTM-Hawkes is made up of four consecutive steps, A) model configuration, B) parameter optimization by minimizing Eq (12), C) finding derivative γ which minimize residual error, and D) forecast event times during the prediction period and draw $\lambda(t)$ (Flow Diagram 1). Those steps and each algorithm are described in Algorithm 1, 2, 3, and 4 as pseudo codes above. In the Algorithm 1, we chose evaluation points by dividing the distance $[t_1, t_n)$ with the size of eP which means at the implementation level, Δt which is $t - t_i$ in the Eq (8), is equal to τ_n. After that, the hyper parameters are defined with the initial values and optimization is proceeded. In the Algorithm 4, by calculating lambda intensity with the results from LSTM prediction $T_i^{'}$, our model finally draws cumulative intensity values for $[t_1, t_n^{'})$

Results and analysis

Data. Criteo, an Ad Tech company, possesses cutting edge user re-targeting technology and has led the development of prediction methods based on machine learning algorithms. They have also notably hosted the Kaggle's CTR forecast competition in 2014 and 2015. We applied the click and conversion data sets provided by the Criteo Lab and compared the results between the the model/algorithm proposed and the HIP (Hawkes Intensity Process algorithm) [21] algorithm. We evaluated the forecasting accuracy of the posterior distribution of a click event from a single advertising network.

Goodness of Fit. Residual analysis [19] is a reliable measurement for the Hawkes model and has been widely used to evaluate the precision of a fit. Let t_i be a point process with intensity $\lambda(t_i)$ whose PDF is Eq (10) and s_i is equal to $\lambda(t_i)$ of Eq (5), then s_i is a unit rate Poisson process transformed by a Hawkes model. Thus, if the model fits well, the transformed process should resemble a unit Poisson process. Also, the residual's inter-event time is supposed to be an independent exponential variable. Therefore, log-log-plot of the residual's inter-event time should be close to the linear line. Fig 3 shows the log-log-plot of the residual inter-event time from the LSTM-Hawkes model and it proves that the shape of the distribution is very close to the linear line. Also, the first quintile of residual's inter-event times are distributed the most. Thus, we can conclude that the LSTM-Hawkes model is an excellent way to determine a precise fit.

Compared Algorithm HIP. The HIP model is a Hawkes-based model that uses Power-Law-Kernel Eq (7). HIP mathematically induces the expectation function $\xi(t)$ of $\lambda(t)$ referring to exogenous events to predict the number of occurrences of an event during the next time unit. In our test, an ad-click is set up as an endogenous event and conversion as an exogenous event.

$$\xi(t) = Expectation\ of\ \lambda(t) \quad where \quad \lambda(t) = \mu s(t) + \sum_{t_i<t} k m^{\beta} (\tau + c)^{-(l+\theta)}. \tag{13}$$

Error Score.

$$s = \begin{cases} \sum_{i=1}^{n} e^{-(d/a_1)} - 1 \quad for\ d < 0 \\ \sum_{i=1}^{n} e^{(d/a_2)} - 1 \quad for\ d \geq 0 \end{cases} \tag{14}$$

Commonly used error measurements such as the Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE) are only designed

to measure the raw residuals. However, different from these methods, scores metric [22] is devised to take the negative error into account, which reduces the error score when the residuals are less than a certain point. Not only does the score metric method discover how many residuals the model produces, but it is also able to discover how well the model fits and predicts. In the experiment, we set up this certain point using the standard deviation of the observed data during the time period for the prediction. In the 1st test section, both a_1 and a_2 are 4.057 and in the 2nd test section both a_1 and a_2 are 7.495. The final score is the sum of each function conditioned on the distance mark.

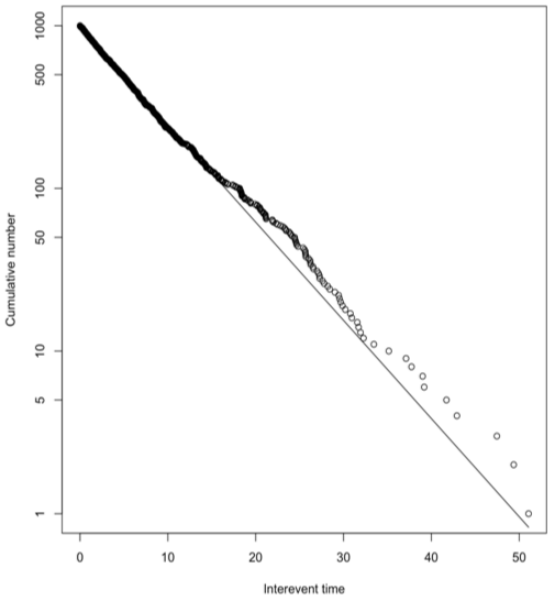
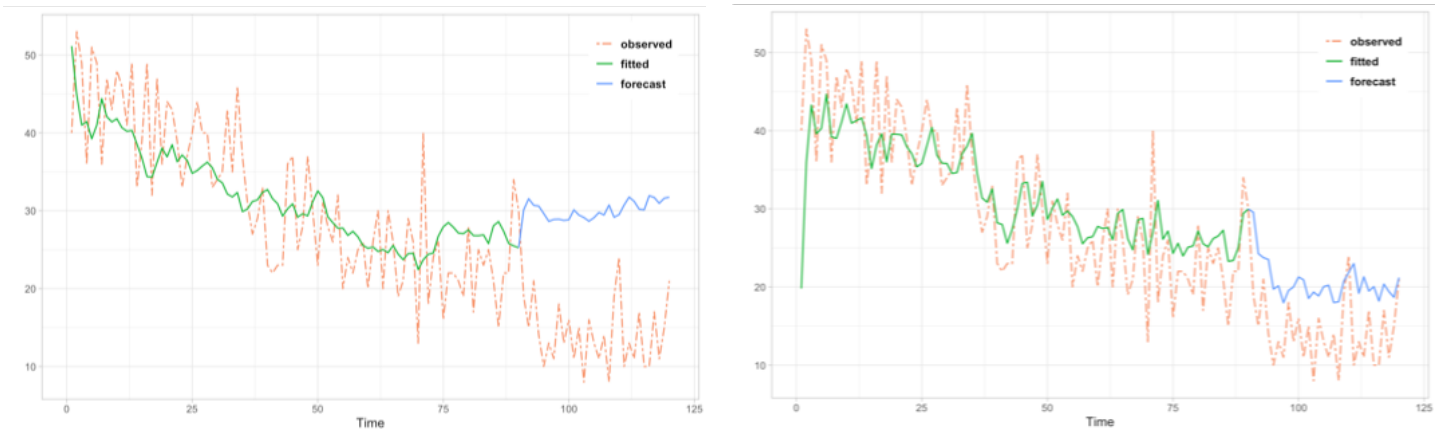


Fig 3. log-log-plot of residual's interevent times 1000 events sampled.

Experimental Results. To asses the comparison between the HIP and LSTM Hawkes models, we selected two different sections where the moving average trend of the click is monotonically decreasing in the first section and increasing in the second section. Test results of the first section are presented in Fig 4, Fig 5, and table 2. Also, test results of the second section are presented in Fig 6, Fig 7, and table 3. For an accurate performance assessment of the HIP model, we chose several correlation coefficients so that the HIP model could show its accuracy in prediction, precision in fitting, as well as it's limitation in both the fitting and prediction. For the accuracy of our assessment we chose four equally dispersed correlation coefficients of 0.2, 0.4, 0.6, and 0.8. The observed click is the endogenous event, and the exogenous event is the observed conversion.

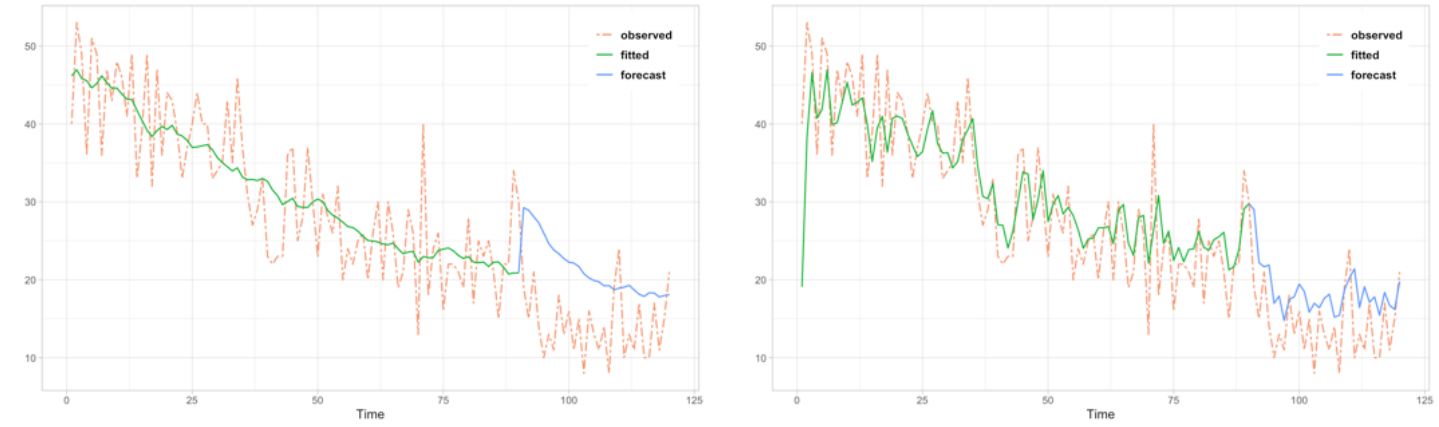
First, in the forecasting test in the first section, it shows that the forecast made using the LSTM-Hawkes model reduces 67.9% of MSE on average and at least 37.8% compared to the HIP model where the correlation coefficient is 0.8. The forecasting test for the second section shows that LSTM-Hawkes reduces 62.3% of MSE on average and at least 44.5% compared to the HIP where the correlation coefficient is 0.6. Fig 4-(a) and Fig 4-(b) also shows the difference between the HIP and Hawkes models where a sudden drop occurs at the beginning part of prediction period. In contrast to HIP that does not follow the moving average due to the low correlation coefficient, LSTM-Hawkes follows the moving average closely, greatly reducing residuals. In fig 6-(a), HIP follows the moving average closely with a high coefficient (0.8) of exogenous

data but still it's MSE is 26 points higher than that of LSTM-Hawkes.



(a) HIP where the coefficient of exogenous event is 0.2 **(b) LSTM-Hawkes with Nelder-Mead**

Fig 4. HIP with coefficient 0.2 shows the most MSE in HIP test in the 1st section (increasing trend) as same as LSTM-Hakwes with SANN does in LSTM-Hawkes Test.



(a) HIP where the coefficient of exogenous event is 0.8 **(b) LSTM-Hawkes with Conjugate Gradient Descent**

Fig 5. As correlation coefficient gets higher, HIP detects the sudden drop better. However since HIP $\xi(t)$ is the expectation value of $\lambda(t)$ with power-law decay, it tends to follow the trend rather than predict actual intensity value each time unit.

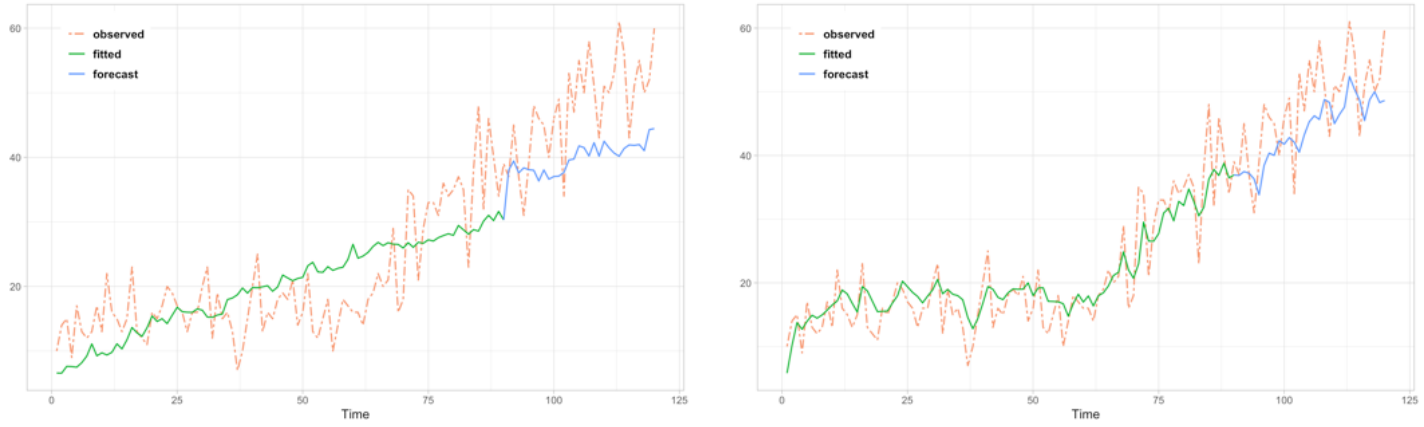
Table 2. MSE and Accuracy

HIP		**LSTM-Hawkes**	
Correlation Coefficient	**MSE**	**Method**	**MSE**
0.20045197016606	272.86493629506	Nelder-Mead	55.3646851328
0.40864585590557	167.08929263969	BFGS with Hessian Matrix	31.8141777151
0.60923578698778	166.47562141766	Conjugate Gradient Descent	44.5680031396
0.80069927612292	75.905022392	SANN	47.1453170336
Average Accuracy: 0.247342286175		Average Accuracy: 0.531401427812	

MSE and Accuracy where the trend of moving average is monotonically decreasing.

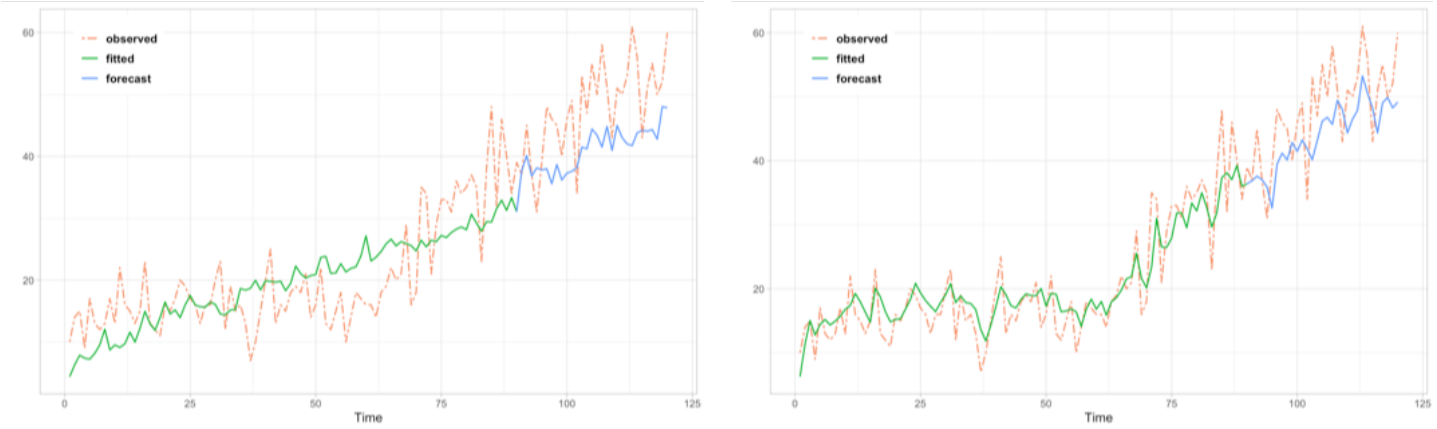
In the second section, although both HIP (coefficient 0.6) and LSTM-Hawkes show great performance in prediction as can be seen in Fig7, between 50 and 70 less-fitted

parts are found with HIP while LSTM-Hawkes shows a stable goodness of fit. This gap is also shown in Fig 6-(a) as well. From the tests, we were able to verify that our proposed model of the LSTM-Hawkes significantly outperforms the HIP model in both the fitting and forecasting criteria.



(a) HIP where the coefficient of exogenous event is 0.8 **(b) LSTM-Hawkes with BFGS**

Fig 6. Rather than predictive period, it needs to compare the observed time period more. Since an under-fitted part is found with HIP result, prediction performance is quite noticeable.



(a) HIP where the coefficient of exogenous event is 0.6 **(b) LSTM-Hawkes with SANN**

Fig 7. The under-fitted part is still found with HIP (the coefficient of exogenous event is 0.6) while both algorithm perform great in prediction resulting 76.9 and 40.7 MSE respectively.

Table 3. MSE and Accuracy

HIP		**LSTM-Hawkes**	
Correlation Coefficient	**MSE**	**Method**	**MSE**
0.20299594548747	801.44661263499	Nelder-Mead	48.0034938190
0.40473704703628	101.32908279598	BFGS with Hessian Matrix	41.3346206943
0.60048848461137	76.966689384563	Conjugate Gradient Descent	42.6831118624
0.79396502231167	84.072450398227	SANN	40.7298560311
Average Accuracy: 0.741641169307		Average Accuracy: 0.882077335865	

MSE and Accuracy where the moving average trend is monotonically increasing.

Pertaining to the compatibility test with various convex optimization algorithms based on both gradient descent method and Newtonian method, we could prove that LSTM-Hawkes has great compatibility with all of the tests. We set up our model for the compatibility test with Nelder-Mead [23], BFGS with Hessian Matrix [24], Conjugate Gradient Descent [25], and Simulated Annealing [26] and have our model optimized by estimating hyper parameters with the convex optimization functions mentioned above. Fig 8 shows learning results based on these optimization functions and we could discover that the full interquartile range (IRQ) of MAE is a lot smaller than that of HIP. Also, since our model is not dependent with exogenous data to refer, the outlier-points cannot be found. It is expected that compatibility with different optimization algorithms will always be guaranteed.

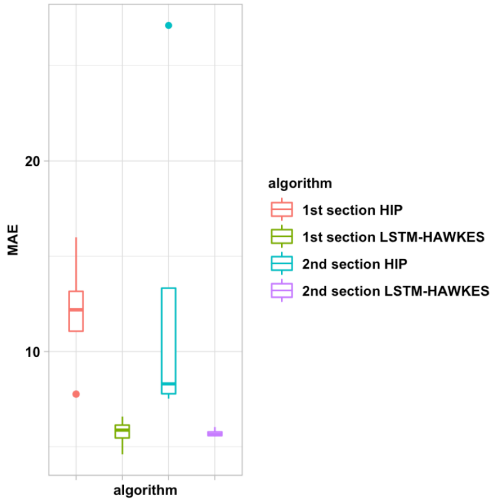


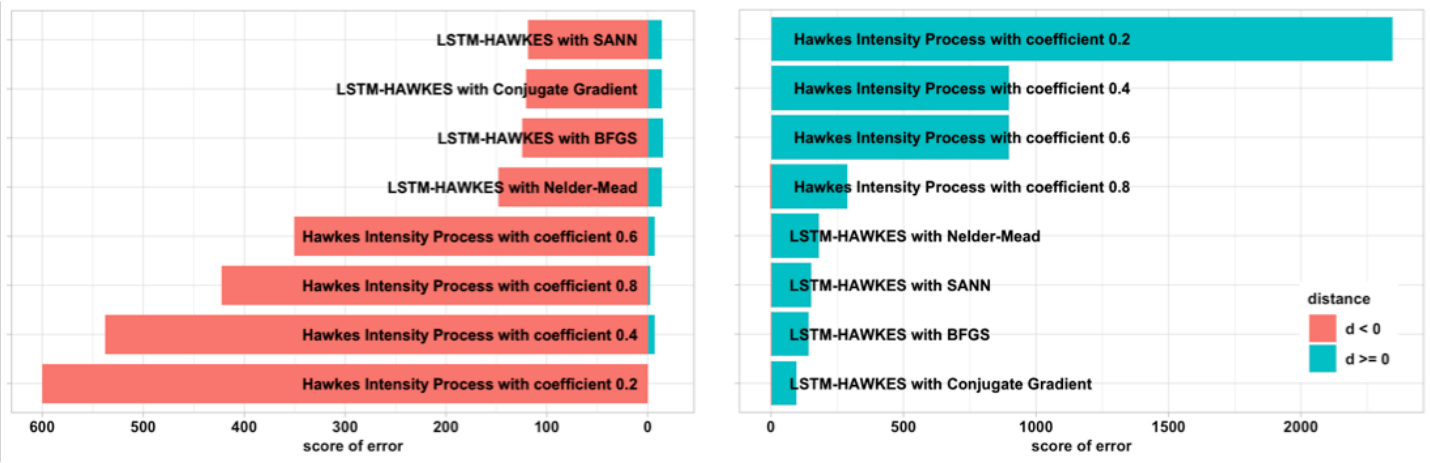
Fig 8. Comparability Test Full quantile range of Mean Absolute Error result of both HIP and LSTM-Hawkes.

Lastly, when we assessed the score metric, Eq (14), of the two models, we were able to see that the LSTM-Hawkes model always obtained a greater negative error score than that of the HIP model. Thus, this always resulted in a lower score metric for the LSTM-Hawkes model, proving it's greater accuracy of prediction and precision of fit.

Conclusion

Conclusion. In 2014, Ian Goodfellow suggested Generative Adversarial Network (GAN) [27]. The combination of generative model and discriminative model was a brilliant approach which maximizes the ability of a pair of discriminative model and generative model. We also aimed to build a hybrid model to apply to forecasting. However, it was practically difficult to apply models such as GAN in Ad Tech due to its high computations. Our Hawkes model is designed as a single-layer perceptron consisted of a much fewer number of parameters than those of feature driven Neural Network while it still predicts/models the posterior distribution accurate and stable. By adapting Hawkes as generative model instead of Neural Network, we can expect easier implementation and computation efficiency. Also, by adapting LSTM as predictive model, higher forecast accuracy is expected to be obtained in comparison to the existing Hawkes Model.

Future Work. These days, to tract more consumers, advertisers strongly focus on



(a) Test Section 2: Increasing Trend **(b) Test Section 1: Decreasing Trend**

Fig 9. Under the circumstance that the moving average of ad clicks slopes downward, the HIP gets a higher error score where the distance mark d in Eq (14 is negative, and vice versa where the moving average of ad clicks slopes upward.

conversion events which actually make sales. Thus, mostly mobile advertising bidding is on a basis of conversion events such as download, purchase or add to cart. However, since conversion event occurs rare, existing models including Hawkes and other Neural Network models have difficulties in predicting its posterior distribution. In the future research, we plan to develop/implement a Conversion Ratio (CVR) prediction model which overcomes data sparsity problem.

References

1. Advertising Expenditure Forecasts. Zenith Media report 2018.
2. Lee, Heejun, and Chang-Hoan Cho. Digital advertising: present and future prospects. International Journal of Advertising (2019): 1-10.
3. Shin HyeRim. Mobile advertising market expenditure exceeded 2 trillion won. Cheil Worldwide 2018.
4. Chowdhury, Humayun Kabir. Consumer attitude toward mobile advertising in an emerging market: An empirical study. International Journal of Mobile Marketing 1.2 (2006).
5. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.
6. Hawkes, Alan G. Spectra of some self-exciting and mutually exciting point processes functions. Biometrika 58.1 (1971): 83-90.
7. Juan, Yuchin. Field-aware factorization machines for CTR prediction. Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.
8. Han, Min Ho. Retargeting advertising product recommending user device and service providing device, advertising product recommending system including the same, control method thereof, and non-transitory computer readable storage

medium having computer program recorded thereon. U.S. Patent Application No. 15/320,632.

9. Cross-platform HTTP debugging proxy server application. https://www.charlesproxy.com/overview/about-charles.

10. Feller, W. (1971) Introduction to Probability Theory and Its Applications, Vol II (2nd edition),Wiley. Section I.3 ISBN 0-471-25709-5.

11. Rizoiu, Marian-Andrei. Expecting to be hip: Hawkes intensity processes for social media popularity.. Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.

12. Kwak, Haewoon. What is Twitter, a social network or a news media?. Proceedings of the 19th international conference on World wide web. AcM, 2010.

13. Ahn, Yong-Yeol. Analysis of topological characteristics of huge online social networking services. Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

14. Ozaki, Tohru. Maximum likelihood estimation of Hawkes' self-exciting point processes. Annals of the Institute of Statistical Mathematics 31.1 (1979): 145-155.

15. Hochreiter, Sepp, and Jürgen Schmidhuber. Long short-term memory. Neural computation 9.8 (1997): 1735-1780.

16. Ogata, Yosihiko. On Lewis' simulation method for point processes. IEEE Transactions on Information Theory 27.1 (1981): 23-31.

17. Rasmussen, Jakob Gulddahl. Temporal point processes: the conditional intensity function. Lecture Notes, Jan (2011).

18. Rubin, Izhak. Regular point processes and their detection. IEEE Transactions on Information Theory 18.5 (1972): 547-557.

19. Lorenzen, F. Analysis of order clustering using high frequency data: A point process approach. Working Paper, 2012.

20. $labs.criteo.com/wp-content/uploads/2014/07/criteo_conversion_logs.tar.gz$

21. Rizoiu, Marian-Andrei, et al. "Expecting to be hip: Hawkes intensity processes for social media popularity." Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017.

22. Saxena, Abhinav, et al. "Damage propagation modeling for aircraft engine run-to-failure simulation." 2008 international conference on prognostics and health management. IEEE, 2008.

23. McKinnon, Ken IM. "Convergence of the Nelder–Mead Simplex Method to a Nonstationary Point." SIAM Journal on Optimization 9.1 (1998): 148-158.

24. Berahas, Albert S., Jorge Nocedal, and Martin Takác. "A multi-batch L-BFGS method for machine learning." Advances in Neural Information Processing Systems. 2016.

25. Hestenes, Magnus Rudolph, and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. Vol. 49. No. 1. Washington, DC: NBS, 1952.

26. Granville, Vincent, Mirko Krivánek, and J-P. Rasson. "Simulated annealing: A proof of convergence." IEEE transactions on pattern analysis and machine intelligence 16.6 (1994): 652-656.

27. Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

28. Improvement of the LPWAN AMI backhaul's latency thanks to reinforcement learning algorithms 2018Rémi Bonnefoi

29. Boubchir, Larbi, Somaya Al-Maadeed, and Ahmed Bouridane. "Undecimated wavelet-based Bayesian denoising in mixed Poisson-Gaussian noise with application on medical and biological images." 2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2014.

30. Pérez, Patrick, Michel Gangnet, and Andrew Blake. "Poisson image editing." ACM Transactions on graphics (TOG) 22.3 (2003): 313-318. The