A Bayesian Multi-Armed Bandit Algorithm for Bid Shading in Online Display Advertising

Mengzhuo Guo[*]
mengzhguo@scu.edu.cn
Sichuan University
Chengdu, Sichuan, China

Wuqi Zhang[*]
mfishzhang@tencent.com
Tencent
Shenzhen, Guangdong
China

Congde Yuan[*]
biglongyuan@tencent.com
Tencent
Shenzhen, Guangdong
China

Binfeng Jia
vincejia@tencent.com
Tencent
Shenzhen, Guangdong
China

Guoqing Song
karrysong@tencent.com
Tencent
Shenzhen, Guangdong
China

Hua Hua
mikehua@tencent.com
Tencent
Shenzhen, Guangdong
China

Shuangyang Wang
feymanwang@tencent.com
Tencent
Shenzhen, Guangdong
China

Qingpeng Zhang
The University of Hong Kong
Hong Kong, China
qpzhang@hku.hk

Abstract

In real-time bidding systems, ad exchanges and supply-side platforms (SSP) are switching from the second-price auction (SPA) to the first-price auction (FPA), where the advertisers should pay what they bid if they win the auction. To avoid overpaying, advertisers are motivated to conceal their truthful evaluations of impression opportunities through *bid shading* methods. However, advertisers are consistently facing a trade-off between the probability and cost-saving of winning, due to the information asymmetry, where advertisers lack knowledge about their competitors' bids in the market. To address this challenge, we propose a **Bayes**ian **M**ulti-**A**rmed **B**andit (BayesMAB) algorithm for bid shading when the *winning price* is unknown to advertisers who lose the impression opportunity. BayesMAB incorporates the mechanism of FPA to infer each price interval's winning rate by progressively updating the market price hidden by SSP. In this way, BayesMAB better approximates the winning rates of price intervals and thus is able to derive the optimal shaded bid that balances the trade-off between the probability and cost-saving of winning the impression opportunity. We conducted large-scale A/B tests on Tencent's online display advertising platform. The cost-per-mile (CPM) and cost-per-action (CPA) decreased by 13.06% and 11.90%, respectively, whereas the return on investment (ROI) increased by 12.31% with only 2.7% sacrifice of the winning rate. We also validated BayesMAB's superior performance in an offline semi-simulated experiment with SPA data sets. BayesMAB has been deployed online and is impacting billions of traffic every day. Codes are available at https://github.com/BayesMAB/BayesMAB.

CCS Concepts

• **Information systems → Computational advertising**; **Display advertising**.

Keywords

Computational advertising, Online display advertising, Real-time bidding, First price auction, Multi-armed bandit

ACM Reference Format:
Mengzhuo Guo, Wuqi Zhang, Congde Yuan, Binfeng Jia, Guoqing Song, Hua Hua, Shuangyang Wang, and Qingpeng Zhang. 2024. A Bayesian Multi-Armed Bandit Algorithm for Bid Shading in Online Display Advertising. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3627673.3680107

1 Introduction

Motivation. The real-time bidding (RTB) paradigm is a major source of revenue in online display advertising [21]. As a dominant form of auction in RTB, the *second-price auction* (SPA) allows advertisers to pay the second-highest bid price in the market [11, 28]. However, many Ad Exchanges (AdX) and Supply-Side Platforms (SSP) began to switch from SPA to generalized *first-price auction* (FPA). In the FPA market, the demand-side platforms (DSP) or advertisers pay what they bid if they win. Advertisers use a strategy called *bid shading* to conceal their true valuation of each impression opportunity while still winning with a lower bid, to avoid intentional overpayment [18].

Challenge. The primary challenge in bid shading pertains to information asymmetry, where advertisers lack knowledge about their competitors' bids in the auction. Specifically, if an advertiser wins an impression opportunity, it may end up overpaying due to the absence of market transparency. Conversely, if the advertiser loses, the content publisher has no obligation to disclose the gap between the advertiser's bid and the *market price*, i.e., the minimum bid required to win the auction. Thus, advertisers face the risk of losing the competition if they underbid while overpaying if they overbid [30]. This work focuses on (1) *incorporating FPA property in optimal bidding algorithm*, and (2) *balancing the trade-off between exploiting the bid with a higher chance of winning and exploring lower bids to reduce costs.*

[*] These authors contributed equally to this research.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680107

In this work, we formulate the bid shading problem into a Bayesian Multi-Armed Bandit (BayesMAB) approach. The proposed BayesMAB accounts for the mechanism of FPA, i.e., when an arm, representing a bid, b wins/loses the impression opportunity, the arms greater/smaller than b will also win/lose. However, bids greater than b should be penalized to avoid overpaying. Therefore, the rewards are related to (a) the gap between the bid and the market price (unknown to the advertisers), and (b) the number of winning impressions of the selected arms. The penalization depends on the gap between the selected bid and the market price. BayesMAB estimates the unknown market price by exploring rather than learning its distribution. The posterior of the winning rates over different arms is learned by progressively updating a Beta distribution in a Bayesian manner, given the results of exploiting the arm with the greatest reward and the deduced market price (Algorithm 1). Algorithm 2 then searches the optimal shaded bid considering each impression opportunity's evaluated value and the learned winning rate of each price interval to balance the probability and cost-saving of winning. We allow the distribution to be non-stationary in online settings. The online results conducted on a real advertising platform showed that the cost-per-mile (CPM) and cost-per-action (CPA) decreased by 13.06% and 11.90%, whereas the return on investment (ROI) increased by 12.31% with only 2.7% sacrifice of the winning rate.

Contribution. First, we propose an algorithm based on MAB for bid shading problems. To the best of our knowledge, we are the first to incorporate the knowledge of FPA into a Bayesian MAB framework. The proposed algorithm better describes the advertisers' searching process for the optimal shaded bid, thus effectively balancing the probability and cost-saving of winning the impression opportunity. Moreover, although the proposed algorithm is deployed online, we propose an intuitive process for using SPA data to help examine offline FPA experiments when the losing impressions' market prices are not revealed to the advertisers. At last, we conducted large-scale A/B experiments to verify the effectiveness of the proposed algorithm. The online results show that three critical business metrics significantly increased without sacrificing a winning rate. The proposed algorithm is now live on Tencent's online bidding platform, impacting billions of traffic daily.

2 Related Work

Bid Shading Methods in Online Display Advertising. Given whether the market price is known, the bid shading methods can be classified into two streams. If the market price is given, the supervised machine learning approaches can be used to predict an optimal factor determining the extent to which the current bid should be shaded so as not to overpay or lose the impression opportunity. Gligorijevic et al. utilized regression-based models to minimize an asymmetric loss between the shaded price and the true market price but only considered low dimensional feature interactions through a factorization machine [10]. Instead of directly predicting a shading factor, Zhou et al. used a deep neural network to model the complex feature interactions, learn the bidding landscape of the market price, and then search for the optimal bid with the maximization of total surplus [30].

The methods in the second stream assume that the market price is hidden, which is often the case, and formulate the bid shading problem into a surplus maximization scheme. In Pan et al., a modified logistic regression model is proposed to predict the profit from the possible shaded price by estimating the corresponding winning rates [18]. Similarly, Zhou et al. assumed the cumulative distribution function of the market price is in a specific shape, used a deep neural network to approximate the distribution, and then proposed a golden search algorithm to detect the optimal bid [30]. Without pre-defining the distribution of the winning rate, Karlsson and Sang used a two-parametric nonlinear function to describe the shading mechanism [14]. By proposing an adaptive online learning scheme, Karlsson and Sang depicted the relationship between the surplus and the optimal parameters [14]. Despite the parametric assumptions for the distribution, Zhang et al. proposed a non-parametric method enabling the exploitation of structures in real FPA scenarios [29]. The authors showed that the proposed non-parametric approaches could capture more surplus than the parametric ones.

There have been some methods for exploring the distribution of the market in the market, although they still need to address the application in bid shading problems. Wu et al. proposed a mixture model combining linear and censored regression models to handle the situation where the advertisers suffered from censoring the winning price [27]. On this basis, a generalized deep learning model is proposed to learn a more flexible and accurate landscape of the winning price under various distributions, and structures in practice [26]. Without the assumption of winning price distribution, Ren et al. combined deep recurrent neural networks and survival analysis to handle the censorship [19]. Although Ren et al. achieved a state-of-the-art performance when facing sophisticated distributions, the splits of price into trivial intervals would weaken model efficiency [16]. To this end, Li et al. devised a novel neighborhood likelihood loss to learn the arbitrary distribution of the winning prices without any pre-assumptions [16].

All the methods mentioned above only considered censorship in FPA but did not account for the mechanism of FPA. On the other hand, the existent addressed the exploitation process of utilizing the best bid price with the highest winning probability but ignored the importance of the exploration process in trying out different bid prices, which are very likely to win with lower costs.

Multi-armed Bandit Algorithms in Bid Shading. MAB is a powerful framework for algorithms that make decisions in an exploration-exploitation fashion over time under uncertainty [5, 8, 23]. MAB algorithms have been applied to online display advertising, including pricing with discounted valuations [17], optimizing bidding strategies under multiple business objectives [9, 13], acquiring customers [22], and setting a reserve price [20]. Unfortunately, few researchers have focused on the exploration-exploitation tradeoff in bid shading. Han et al. theoretically provided a regret bound in an online stochastic multi-armed bandits (MAB) algorithm considering the FPA with censored feedback [12]. The most relevant research is Tilli and Espinosa-Leal, where the authors simply compared several existing MAB algorithms for the bid shading problem under the FPA context [25]. Our work is different from Tilli and Espinosa-Leal's since we propose a new Bayesian MAB algorithm incorporating the properties of FPA [25].

3 Method

3.1 The Proposed Bayesian Multi-Armed Bandit for Bid Shading

Defining arms. Following the idea in Ren et al.[19], we divide the continuous bid prices into finite K intervals, i.e., $b^0 < \cdots b^k \cdots < b^K$, $k \in \mathcal{K} = \{0, 1, 2, \ldots, K\}$, where b^0 and b^K are the smallest and largest bid prices in the whole price space. Hence, there are K arms defined as $\mathcal{B} = \{\frac{b^0+b^1}{2}, \ldots, \frac{b^{K-1}+b^K}{2}\}$, which is the average of the bounds of each interval. The expected revenue $b_t^k \in \mathcal{B}$ with selecting k-th arm in round $t \in \{1, \ldots, T\}$ is:

$$\mathbb{E}_{\mathcal{D}}\{R_t(b_t^k)\} = \mathbb{P}\{P_t \leq b_t^k\} \cdot \mathbb{E}_{\mathcal{D}}\{r_t(b_t^k) | P_t \leq b_t^k\} \tag{1}$$

where $\mathbb{E}_{\mathcal{D}}\{X\}$ is the expectation of variable X with respect to the distribution $\mathcal{D}$, $r_t(b_t^k)$ is a reward function, P_t is the market price, and $\mathbb{P}(P_t \leq b_t^k)$ is the winning probability if we submit a bid at a price b_t^k. Note that although $\mathcal{D}$ can be considered stationary in a very short period for simplicity, we allow $\mathcal{D}$ to be non-stationary along with time in the online settings. Each arm in $\mathcal{B}$ is a bid that an advertiser is willing to pay for an impression opportunity. Since the market price P_t is not given in FPA, the biggest challenge is that we cannot know the outcome of choosing b_t^k in round t unless we conduct an online test using b_t^k.

To validate the proposed methodology offline, we establish the outcome as a random variable that conforms to a Bernoulli distribution, denoted as $S_t(b_t^k) \sim \mathcal{BN}(\theta_t^k)$. Specifically, $S_t(b_t^k) = \mathbb{I}\{P_t \leq b_t^k\}$ and θ_t^k represents the probability of success when bidding at price b_t^k. It is important to note that within the context of FPA, θ_t^k cannot be known a priori due to the unavailability of market prices to advertisers. Consequently, we assume that θ_t^k follows a Beta distribution $\mathcal{B}(\alpha_{k,t}, \beta_{k,t})$, where $\alpha_{k,t}$ and $\beta_{k,t}$ represent the cumulative number of successful and unsuccessful impression opportunities, respectively, if we bid with the k-th arm in round t. Both $\alpha_{k,t}$ and $\beta_{k,t}$ will be adjusted based on the feedback from the algorithm.

Updating posterior distribution. Denote the distribution of arm k at the end of round $t-1$ by $\mathcal{B}(\alpha_{k,t-1}, \beta_{k,t-1})$. Assume in round t, we select the arm $k^* \in \mathcal{K}$ and we can observe the outcome $S_t(b_t^{k^*}) = s_t^{k^*}$ either by investigating real bidding log data or sampling $s_t^{k^*} \sim \mathcal{BN}(\theta_{t-1}^k)$. If $b_t^{k^*}$ can win the impression opportunity, i.e., $s_t^{k^*} = 1$, determine $s_t^k, k \in \{1, \ldots, k^*-1\}$ in the same way of producing $s_t^{k^*}$. Define $k_t^-(k^*)$ by

$$k_t^-(k^*) = \max\{k \in \{1, \ldots, k^*-1\} | s_t^k = 0, s_t^{k^*} = 1\} \tag{2}$$

then $k_t^-(k^*)+1$ is the index of the arm with the lowest price that *can* win the bid. In this case, since the real market price P_t is not known a priori due to information deficiency, we estimate the market price with $\hat{P}_t = b_t^{k_t^-(k^*)+1}$. Similarly, if $b_t^{k^*}$ loses, i.e., $s_t^{k^*} = 0$, determine $s_t^k, k \in \{k^*+1, \ldots, K\}$. Define $k_t^+(k^*)$ by

$$k_t^+(k^*) = \min\{k \in \{k^*+1, k^*+2, \ldots, K\} | s_t^k = 1, s_t^{k^*} = 0\} \tag{3}$$

then $k_t^+(k^*)$ is the index of the arm with the lowest price that *can* win the bid, in which case the market price $\hat{P}_t$ is estimated with $b_t^{k_t^+(k^*)}$. Simply put, we can always win at the minimum cost if we bid $b_t^{k_t^+(k^*)}$ (when the selected arm loses) or $b_t^{k_t^-(k^*)+1}$ (when the selected arm wins).

We consider two procedures for updating Beta distribution in round t according to the properties of FPA: (1) If $s_t^{k^*} = 1$, then all the arms with $j \in \{k_t^-(k^*)+1, \ldots, k^*-1, k^*, k^*+1, \ldots, K\}$ can win the impression opportunity so that the posterior distribution of arm j should be $\mathcal{B}(\alpha_{k,t-1}+1, \beta_{k,t-1})$; (2) If $s_t^{k^*} = 0$, then all the arms with $j \in \{1, \ldots, k^*, \ldots, k_t^+(k^*)-1\}$ will lose the impression opportunity so that the posterior distribution of arm j should be $\mathcal{B}(\alpha_{k,t-1}, \beta_{k,t-1}+1)$.

Defining reward function. We here utilize a *pseudo weighted expected returns* $m(k,t)$ to represent $\mathbb{E}_{\mathcal{D}}\{r_t(b)|P_t \leq b_t^k\}$ in Eq.(1), accounting for the FPA's mechanism. The reward function considers two parts: (1) the relative gap between $b_t^k, k \in \mathcal{K}$ and the market price $\hat{P}_t$; (2) the number of winning impression opportunities of arm k until t rounds:

$$m(k,t) = \begin{cases} \frac{w_0^k \cdot \alpha_{k,0}}{\beta_{k,0}}, & \text{if } t = 0 \\ \frac{\sum_{i=1}^t s_i^k \cdot w_i^k}{t}, & \text{otherwise} \end{cases} \tag{4}$$

where $w_i^k = \frac{1}{\exp\{|b_i^k - \hat{P}_i|\}}, i \in \{0, \ldots, t\}$. We initialize $\hat{P}_0$ with the median price of the historical *winning* impressions. As the iteration goes by, the market price is deduced at t-th round when $t > 0$, i.e., $\hat{P}_t = b_t^{k_t^-(k^*)+1}$ if $s_t^{k^*} = 1$, and $\hat{P}_t = b_t^{k_t^+(k^*)}$, otherwise.

Eq.(4) considers the number of winning impressions and the difference between the market price and the k-th arm. In FPA, if arm k wins but there is another bid b_t^j with $\hat{P}_t \leq b_t^j < b_t^k$, we search for $\hat{P}_t$ and lower the bid to save costs. The reward should be penalized if overpaying occurs, with the degree of penalization (w_t^k) being related to the distance between the selected arm and $\hat{P}_t$. Unlike in FPA, in SPA, all bids greater than $\hat{P}_t$ can win the impression opportunity without penalty, regardless of their distance from the market price.

Selecting arm. To decide which arm should be selected, we provide an index based on the UCB-like methods. Let $\mathcal{T}_{k,t} \subseteq \{1, \ldots, t\}$ denote the indices of the rounds that arm k is played as k^* until round t. In round $t+1$, we select arm k^* such that

$$k^* = \arg\max_{k \in \mathcal{K}} \{I(k,t) \cdot m(k,t) + \eta(k,t)\} \tag{5}$$

where $I(k,t)$ considers the mean and variance of the posterior distribution of arm k [20]:

$$I(k,t) = \frac{\alpha_{k,t}}{\alpha_{k,t} + \beta_{k,t}} + \frac{\alpha_{k,t} \cdot \beta_{k,t}}{(\alpha_{k,t} + \beta_{k,t})^2 \cdot (\alpha_{k,t} + \beta_{k,t} + 1)} \tag{6}$$

$$\eta(k,t) = \sqrt{\frac{2 \log \sum_{i=1}^K |\mathcal{T}_{k,t}|}{|\mathcal{T}_{k,t}|}} \tag{7}$$

where $|\mathcal{T}_{k,t}|$ is the number of k-th arm being selected up to t-th round. In Eq.(6), the mean of Beta distribution is subject to an increase (first term in $I(k,t)$), while the variance is subject to a decrease (second term in $I(k,t)$) as a result of the repeated play of a given arm k. Consequently, the value of $\eta(k,t)$ in Eq.(7) will be comparatively lower than that of $\eta(j,t)$ for $j \neq k$. To sum up, the mean of $I(k,t)$ in Eq.(6) signifies the exploitation strategy of selecting arms with higher winning probabilities, while the variance component and $\eta(k,t)$ pertain to the exploration of arms with slightly lower probabilities of winning, which can nevertheless be cost-effective in the event of a win.

3.2 Algorithms and Online Deployment

Algorithm 1 presents the pseudo-code for the proposed BayesMAB. The codes in lines 7 and 10 in Algorithm 1 describe the searching process for the market price considering the mechanism of FPA. In particular, if the current arm can win the impression opportunity, so do all arms with greater bids. Similarly, if a selected arm loses, the smaller bids cannot prevail over the other bidders in the market. In practice, BayesMAB is deployed online and uses the real-time feedback - observation of $s_t^{k^*}$ - to update posterior. In the offline setting, since we cannot directly observe the results $s_t^{k^*}$ of using arm k^* in each round due to information deficiency, we draw $s_t^{k^*} \sim \mathcal{BN}(\theta_t^{k^*})$, where $\theta_t^{k^*} \sim \mathcal{B}(\alpha_{k^*,t-1}, \beta_{k^*,t-1})$, to help update the posterior [20].

Algorithm 1 The Bayesian multi-armed bandit algorithm for bid shading

```
Input: Number of arms $K$, set of bids $\mathcal{B}$, set of arms $\mathcal{K}$, parameter
    of Beta distribution $\{(\alpha_{k,0}, \beta_{k,0})\}_{k\in\mathcal{K}}$, number of rounds $T$.
Output: Winning rate $wr^k$ and expected reward $m(k,T)$ for each
    arm $k$ at the end of round $T$.
1:  Initialize $m(k,0)$ and $I(k,0)$, for $k \in \mathcal{K}$.
2:  while $t \le T$ do
3:      Set $t = t+1$ and select arm $k^* = \arg\max_{k\in\mathcal{K}}\{I(k,t-1) \cdot m(k,t-1) + \eta(k,t-1)\}$.
4:      Observe $s_t^{k^*}$.
5:      if $s_t^{k^*} = 1$ then
6:          Determine $s_t^j, \forall j \in \{1,\dots,k^*-1\}$ and $k_t^-(k^*)$ by Eq.(2).
7:          Update posterior $\alpha_{k,t} = \alpha_{k,t-1}+1, k \in \{k_t^-(k^*)+1,\dots,K\}$
            and $\beta_{k,t} = \beta_{k,t-1}+1, k \in \{1,\dots,k_t^-(k^*)\}$.
8:      else
9:          Determine $s_t^j, \forall j \in \{k^*+1,\dots,K\}$ and determine $k_t^+(k^*)$
            by Eq.(3).
10:         Update posterior $\beta_{k,t} = \beta_{k,t-1}+1, k \in \{1,\dots,k_t^+(k^*)-1\}$
            and $\alpha_{k,t} = \alpha_{k,t-1}+1, k \in \{k_t^+(k^*),\dots,K\}$.
11:     end if
12:     Update $m(k,t)$, $I(k,t)$ and $\eta(k,t)$ according to Eq.(4), Eq.(6)
        and Eq.(7), respectively.
13: end while
```

Algorithm 1 approximates the winning rate for each price interval in the *perfect competition market* where a sufficient number of buyers and sellers participate in the transactions as price takers [2]. It indicates that the market itself determines the winning rate and is not influenced by any specific DSP or SSP. We do not account for the value of each traffic in Algorithm 1 since it only affects our willingness to bid higher but does not impact the winning rate in the market. Consequently, the bid shading problem is usually formulated as a two-stage process where we first evaluate the winning rate of each price interval and then provide an optimal shading factor considering each traffic's value [18, 30]. Algorithm 2 uses the cumulative numbers of winning and losing, i.e., $\alpha_{k,T}$ and $\beta_{k,T}$, at the end of Algorithm 1, to produce the winning rate of each arm. The optimal bid is obtained by maximizing the total surplus.

Algorithm 2 Online algorithm for searching optimal shaded bid.

```
Input: Winning rate $wr^k$ from Algorithm 1. Impression
    opportunity $i \in \mathcal{I}$ and its value $v_i$.
Output: Optimal shaded bid $b_i^*$ for each impression opportunity.
Get $wr^k$ for each arm by $wr^k = \frac{\alpha_{k,T}-\alpha_{k,0}}{\alpha_{k,T}-\alpha_{k,0}+\beta_{k,T}-\beta_{k,0}}$.
Get optimal bid $b_i^* = \arg\max_{b_i^k \in \{k\in\mathcal{K}|0\le b_i^k \le v_i\}} wr^k(v_i - b_i^k)$.
Collect results and update $\alpha_{k,0}, \beta_{k,0}$ for the next period.
```

For non-stationary online settings, we divide each day into several periods. For n-th one-hour time window $n \in \{1,\dots,24\}$, we apply Algorithm 1 to pre-train the winning rate, i.e., and then use Algorithm 2 to adjust our basic bid price. In the next period, $\alpha_{k,0}^{n+1} = \kappa\alpha_{k,T}^n$ and $\beta_{k,0}^{n+1} = \kappa\beta_{k,T}^n, 0 < \kappa < 1$, indicating the update of the posterior distribution is discounted along with time, which describes a non-stationary online environment.

Time complexity. The most time-consuming part of Algorithm 1 is the search for the optimal arm and posterior update executed in every round, which renders the time complexity of Algorithm 1 to be $O(2K * T)$. This complexity scales linearly with the number of predefined arms and rounds. Algorithm 2 is an optimization problem with $O(K)$ complexity.

4 Semi-simulated Offline Experiments

The offline experimental results examine three problems: (1) How to determine the parameters in BayesMAB? (2) What is BayesMAB's performance compared to baselines? (3) How important is the exploration process in BayesMAB?

4.1 Data Overview and Experimental Setting

Given our lack of knowledge regarding the true market price in the FPA competition, we propose a semi-simulated approach that utilizes data sets from the SPA market as well as simulated bidding results. Specifically, we collected bidding logs for two advertisement placements from October 17th to October 23rd, 2022, in the SPA market. These logs contain N pairs of prices $\{(ecpm_t, p_t)\}_{t=1}^N$, where $ecpm_t$ represents the evaluation of the value, v_i, and p_t represents the second-highest bid - the winning price - in the market for the t-th impression opportunity. It should be noted that $ecpm_t > p_t > 0$ if we win, and $p_t = 0$ if we lose. Here, the term "placement" refers to a specific location on a content publisher's platform, such as a banner on a webpage or a splash ad on a mobile app at launch.

We use the time-based nested cross-validation approach [24]. Specifically, we use the first day's data to generate the validation set and the next two consecutive days' data as the training and testing sets. For each training-testing data set, we repeat 10 trials, therefore, we conduct 50 experiments for each method and report the average results. The equal frequency binning technique is used to construct $K-1$ sub-intervals so that the number of training data assigned to each of them is the same, i.e., N/K. If N/K is not an integer, we balance the distribution manually so that the numbers of data assigned to different arms differ by, at most, one. The *validation set* is then used to provide prior information for $\alpha_{k,0}$ and $\beta_{k,0}$. At last, Algorithm 2 determines the total saved cost of the winning impressions recorded in the *testing set* given the winning rate of each arm from Algorithm 1.

Placement	Type	#Avg.
A	Banner Ad	279,978
B	Interstitial Ad	290,039

Table 1: The advertisement type of each Placement and the average number of data in the logs. Due to data privacy protocol, we cannot report the real Placement IDs. Both Placements are located on the same mobile app. The banner ad appears above the app's content and stays there until the user hits the "Close" button. The interstitial ad, built into the app script, is displayed when navigating between screens.

Note that there are two conditions when observing t-th data $s_t^{k^*}$ in Algorithm 1: (1) When $p_t \neq 0$, i.e., we have won t-th impression opportunity, if $b_t^{k^*} \geq p_t$, it means that the selected arm can also win so that $s_t^{k^*} = 1$. If $b_t^{k^*} < p_t$, $s_t^{k^*} = 0$. (2) When $p_t = 0$, i.e., we have lost t-th impression opportunity by providing $ecpm_t$, if $b_t^{k^*} \leq ecpm_t$, it means that the selected arm provides a bid smaller than $ecpm_t$ so that we cannot win this time either, i.e., $s_t^{k^*} = 0$. If $b_t^{k^*} > ecpm_t$, we still do not know whether the current optimal bid is greater than the highest market price; therefore, we draw $s_t^{k^*}$ as stated in Algorithm 1. The results of the selected bids are either determined by comparing them to the true winning impressions' prices or by drawing from a distribution that is learned under the FPA's mechanism.

Metrics. The performance of offline experiments can be evaluated from *cost per mile* (*cpm*), winning ratio (*ratio*), and saved costs (*surplus*). Since we do not know the actual winning prices of the impressions that we lose in SPA data sets, we only compare the shaded bids to the winning ones (those $p_i \neq 0$), i.e., the shaded bid wins if $b^* \geq p_i$ and loses otherwise. Denote the index of data in the testing data set by $\mathcal{I}_{test}$, $cpm = \frac{\sum_{i \in \mathcal{I}_w} b_i^*}{|\mathcal{I}_w|}$, where $\mathcal{I}_w = \{i \in \mathcal{I}_{test} | b_i^* \geq p_i, p_i \neq 0\}$. $ratio = \frac{|\mathcal{I}_w|}{|\mathcal{I}_l| + |\mathcal{I}_w|}$, where $\mathcal{I}_l = \{i \in \mathcal{I}_{test} | b_i^* < p_i, p_i \neq 0\}$, and $surplus = \sum_{i \in \mathcal{I}_w}(ecpm_i - b_i^*)$. An effective bid shading algorithm should prioritize maximizing its surplus. Simultaneously, while reducing the initial bids and lowering the *cpm*, the *ratio* should remain comparable to avoid squandering impression opportunities. For data privacy, *surplus* is reported by the number of transformed unit values in the following sections.

4.2 Parameter Selection

The critical parameters in the proposed BayesMAB are the number of arms K and the weight function w_t^k in Eq.(4). Table 2 presents *surplus* of different settings.

Our empirical results indicate that by setting $K = 30$ and using an exponential function for w_t^k, the highest level of *surplus* can be achieved. The exponential form of w_t^k serves as a non-linear penalty on rewards when the chosen arms differ significantly from the inferred market price. Selecting too many arms results in additional exploration spaces, while choosing too few limits the selection of optimal shaded bids, thereby reducing the flexibility in adjusting the original bids in Algorithm 2.

K	Types of w_t^k	Placements A/B
30	Exponential	**4.19**(±0.007)/**8.19**(±0.013)
30	Linear	4.10(±0.011)/8.02(±0.008)
30	Constant	4.15(±0.014)/8.16(±0.012)
20	Exponential	4.15(±0.011)/8.03(±0.014)
50	Exponential	4.12(±0.009)/8.09(±0.008)
70	Exponential	3.90(±0.010)/7.74(±0.010)
100	Exponential	3.66(±0.010)/6.62(±0.007)

Table 2: Results of average $surplus(\pm std)$ applying the shaded bids in the testing data set concerning different K and w_t^k. The best results are marked in bold. Exponential $w_t^k = 1/\exp\{|b_t^k - \hat{P}_t|\}$, linear $w_t^k = 1 - |b_t^k - \hat{P}_t|$, and constant $w_t^k = 1$.

From the perspective of *cpm*, Figures 1a and 1b show that more arms can help lower the average cost per impression since the searching region for the optimal is larger. However, metric *ratio* declines as long as increasing the number of arms on both Placements (see Figures 1c and 1d). In practice, we suggest starting with a small number of arms and iteratively increasing it until the *surplus* is satisfied while both *cpm* and *ratio* are acceptable.

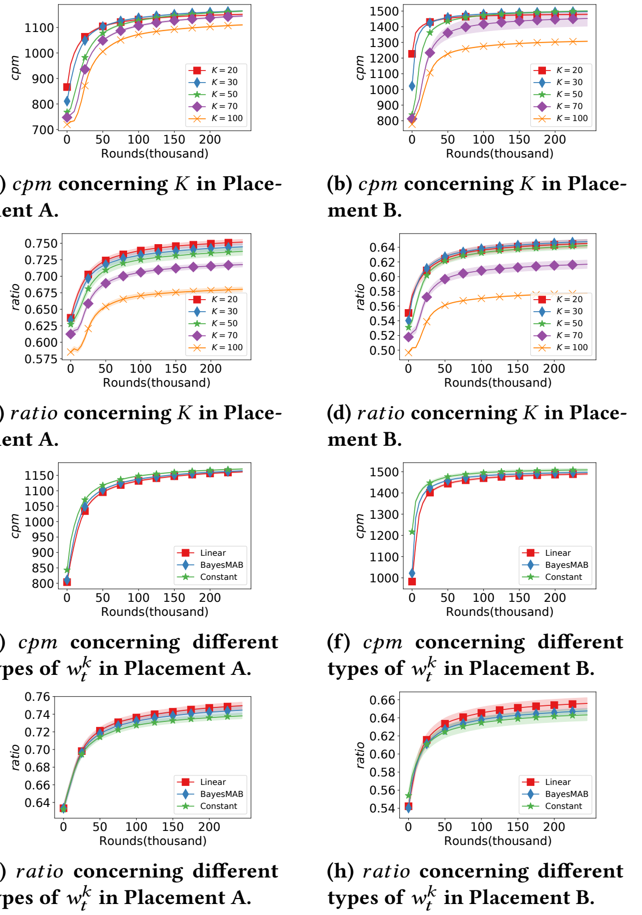
To get a more intuitive understanding of how the FPA mechanism impacts the algorithm, we also compared the performance of different types of w_t^k, as shown in Figures 1e to 1h. The constant $w_t^k = 1$ setting, describing the SPA's mechanism, under-performs the linear and exponential settings since it considers all price intervals equally no matter how far they are from the deduced market price, which ignores the FPA mechanism. Though the linear setting performs the best on Placement A in terms of *ratio*, its performance of *surplus* is smaller than the exponential setting because the latter type penalizes harder on the larger gap between the selected arms and the deduced market price.

4.3 Compared to Baselines

This section compares the proposed algorithm with the baselines, including classic MAB methods, e.g., ϵ-greedy Algorithm (EG) [6], Upper Confidence Bound (UCB1) [4], Upper Confidence Bound Variant (UCB2), Mini-max Optimal Strategy in the Stochastic case (MOSS) [3], Thompson Sampling (TS) [1], Bayesian UCB (BayesUCB) [15], and deep learning-based methods for bid shading, such as Deep Landscape Forecast (DLF) [19] and Efficient Deep Distribution Network (EDDN) [30].

We conduct the same offline experimental settings for MAB algorithms. EG is modified in Algorithm 1 by comparing a random value to $\epsilon = 0.01$. Algorithm 1 only considers the exploration process when the random value is greater than ϵ. UCB1 is tuned according to Theorem 1 in Auer et al.[4]. UCB2 and MOSS are tuned according to Theorem 3 in Degenne and Perchet[7]. DLF and EDDN are first tuned as stated in Ren et al.[19] and Zhou et al.[30], and then produce the winning rate as the input of Algorithm 2.

Table 3 presents the results indicating that the proposed BayesMAB outperforms the baseline algorithms. The difference in *surplus* is



(a) *cpm* **concerning** *K* **in Placement A.**

(b) *cpm* **concerning** *K* **in Placement B.**

(c) *ratio* **concerning** *K* **in Placement A.**

(d) *ratio* **concerning** *K* **in Placement B.**

(e) *cpm* **concerning different types of** w_t^k **in Placement A.**

(f) *cpm* **concerning different types of** w_t^k **in Placement B.**

(g) *ratio* **concerning different types of** w_t^k **in Placement A.**

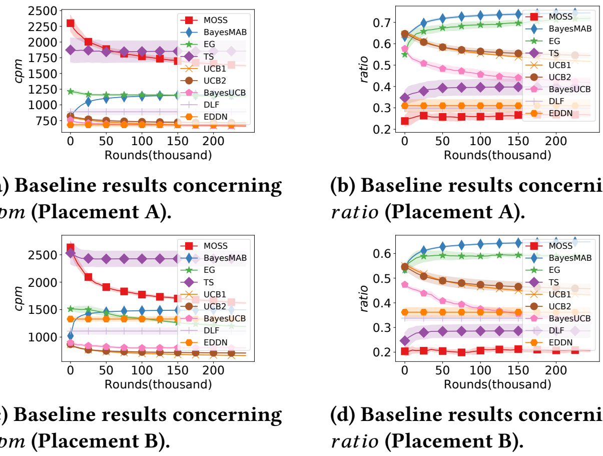
(h) *ratio* **concerning different types of** w_t^k **in Placement B.**

Figure 1: The results of *cpm* **and** *ratio* **with respect to** *K* **and** w_t^k **of Placements A and B. The x-axis is the number of rounds (thousand). The curves and shaded region are the mean and one standard deviation over 50 trials, respectively.**

Methods	Placements A/B
BayesMAB	**4.19**(±0.007)/**8.19**(±0.013)
MOSS	1.92(±0.021)/3.93(±0.048)
EG	3.82(±0.042)/6.18(±0.050)
TS	2.89(±0.264)/6.93(±0.096)
UCB1	1.63(±0.006)/2.49(±0.003)
UCB2	1.68(±0.056)/2.46(±0.057)
BayesUCB	2.47(±0.093)/4.74(±0.065)
DLF	1.37(±0.541)/2.03(±0.872)
EDDN	1.78(±0.620)/3.91(±1.074)

Table 3: Results of *surplus* **compared to baselines.**

at most 9.69% for Placement A and 25.4% for Placement B, as compared to the second-best algorithms. Similarly, although UCB-like methods account for exploring different price intervals, and deep learning-based methods (DLF and EDDN) are proficient in approximating the landscape distribution, they fail to update the rewards in the context of FPA, leading to a significant gap in *surplus* when compared to BayesMAB.



(a) Baseline results concerning *cpm* **(Placement A).**

(b) Baseline results concerning *ratio* **(Placement A).**

(c) Baseline results concerning *cpm* **(Placement B).**

(d) Baseline results concerning *ratio* **(Placement B).**

Figure 2: Results of *cpm* **and** *ratio* **compared to baselines. DLF and EDDN are not iterative algorithms, leading to straight lines.**

Figure 2 presents analogous conclusions. The efficacy of various baselines in terms of rounds is demonstrated by the performance of *cpm* and *ratio*. Despite not having the lowest *cpm* in two Placements, BayesMAB achieves the best *ratio*, indicating that it has balanced the trade-off between cost savings and competitive bids. The *cpm* and *ratio* of UCB-like methods and deep learning-based methods are limited due to excessive shading, resulting in the loss of most impressions. In contrast, TS provides a comparable *ratio* but the worst *cpm* due to insufficient shading. EG obtains similar results to the proposed BayesMAB, as they share a very similar exploration process, and the only difference is the condition for exploration. This suggests that in addition to saving more costs, the design of BayesMAB is more efficient in balancing the trade-off between exploration (lowering *cpm*) and exploitation (maintaining higher *ratio*) than the baselines.

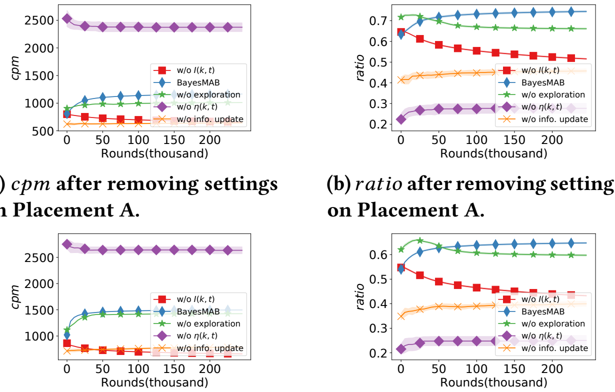
4.4 Ablation Study

This section reports the ablation studies to investigate the importance of the exploration parts when selecting arms and the information update process in Algorithm 1. The results of *surplus* are presented in Table 4.

Remove	Placements A/B
-	**4.19**(±0.007)/**8.19**(±0.013)
$I(k,t)$	1.63(±0.007)/2.49(±0.013)
$\eta(k,t)$	2.53(±0.090)/6.78(±0.0353)
Exploration	3.67(±0.032)/7.91(±0.030)
Info. update	2.35(±0.082)/5.17(±0.073)

Table 4: Results of *surplus* **when removing different terms for describing the mechanism of FPA.**

Our analysis reveals that removal of any term from Eq.(5) leads to a significant decrease in *surplus*, with total reductions of 61.10%/69.60% (when $I(k,t)$ is removed), 39.62%/17.22% (when $\eta(k,t)$ is removed), and 12.41%/3.42% (when both $I(k,t)$ and $\eta(k,t)$ are removed) observed for BayesMAB. These results underscore the crucial role of the exploration process that accounts for the knowledge of the FPA mechanism in enhancing cost-saving capacity. Moreover, we find that the information update process described in Algorithm 1 is equally critical. Eliminating this process results in substantial losses of 43.91% (on Placement A) and 36.87% (on Placement B), indicating that it enables the updating of $\alpha_{k,t}$ and $\beta_{k,t}$ in accordance with the FPA mechanism's rules, which state that if a bid can win an impression, all greater bids can also win.



(a) ***cpm*** **after removing settings on Placement A.**

(b) ***ratio*** **after removing settings on Placement A.**

(c) ***cpm*** **after removing settings on Placement B.**

(d) ***ratio*** **after removing settings on Placement B.**

Figure 3: Results of examining the importance of exploration by blocking terms in Eq.(5).

Furthermore, the outcomes of the convergence of *cpm* and *ratio* are presented in Figure 3. The observed patterns provide further evidence that the integration of the FPA mechanism into the exploration-exploitation process is advantageous for the proposed BayesMAB algorithm, as indicated by the highest *ratio* and satisfactory *cpm*.

5 Online Experiments and Conclusions

We conducted meticulously designed A/B test experiments on an authentic online display advertising platform to establish the effectiveness of the proposed BayesMAB as measured by key performance indicators. Return on Investment (*ROI*) gauges the algorithm's ability to generate profits for advertisers. *CPM* and cost per action (*CPA*) reflect the average cost of securing an impression and acquiring a new user, respectively.

Indicator	Improvement	Indicator	Improvement
ROI	+12.31%***	*CPA*	-11.90%***
CPM	-13.06%***	*winning rate*	-2.7%*

Table 5: Relative improvement over the control group in terms of four business indicators. Results with ***,**,* are statistically different from the control group at $\alpha = 0.01, 0.05, 0.1$.

The BayesMAB's impact was observed on three content publishers (Media) that generate billions of daily traffic. The experimental design ensured that the treatment and control groups were equipped with the same budget and campaign settings, with the only difference being the inclusion of the shading factor in the final bid for the treatment group. The experiments were conducted over one week, and the results, showing the relative improvement of the treatment group over the control group, are presented in Table 5. Specifically, the *ROI* has increased by 12.31%, while *CPA*/*CPM* has decreased by 11.90%/13.06%. Lowering the bids has helped advertisers acquire additional impression opportunities and revenue. Surprisingly, the *winning rate* has not been severely affected, decreasing by only 2.7%, which is an acceptable loss of impression opportunities.

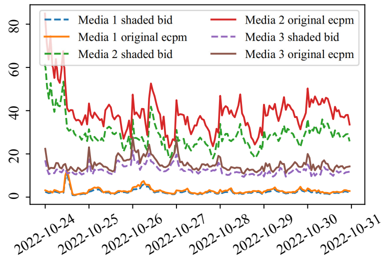


Figure 4: Tendency of average winning prices (shaded bids in FPA) and original bids in the online experiments regarding different media in each hour. The y-axis is the shaded ***ecpm*** **after transformation.**

We graphed the average shaded and original bids for each hour. The results show that BayesMAB proportionally decreases the original bids, ensuring that the shaded bid is competitive in the market while being lower than each impression opportunity's evaluated value. The average shaded bids exhibit a stable pattern that does not dramatically decrease, with the most significant gap between shaded and original bids being smaller than 30%. This observation demonstrates the proposed algorithm's online efficiency and robustness, which now impacts billions of traffic daily.

6 Conclusion

This study proposes a BayesMAB algorithm for bid shading problems, where the bidders are motivated to hide their truthful evaluations of impression opportunities. The bidders face a trade-off between the probability and cost-saving of winning the impression opportunity at a lower price. To this end, BayesMAB incorporates the mechanism of FPA to infer each price interval's winning rate by progressively updating the market price. The offline experiments demonstrated the efficacy of the proposed BayesMAB. Large-scale A/B tests showed that *ROI* increases by 12.31% whereas *CPA*/*CPM* decreases by 11.90%/13.06% without sacrificing too much *winning rate*. BayesMAB is deployed on a real DSP.

Acknowledgments

This work was supported by the Department of Science and Technology of Sichuan Province (#2023YFS0317), China Postdoctoral Science Foundation (#2024M752282), and National Natural Science Foundation of China (#72371176). We thank the reviewers and editors for their constructive comments.

References

[1] Shipra Agrawal and Navin Goyal. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 39–1.

[2] Kenneth J Arrow and Gerard Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society* (1954), 265–290.

[3] Jean-Yves Audibert and Sébastien Bubeck. 2010. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research* 11 (2010), 2785–2836.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[5] Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. 2020. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems* 33 (2020), 1713–1723.

[6] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[7] Rémy Degenne and Vianney Perchet. 2016. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*. PMLR, 1587–1595.

[8] Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. 2021. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems* 34 (2021), 1939–1951.

[9] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. 2013. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[10] Djordje Gligorijevic, Tian Zhou, Bharatbhushan Shetty, Brendan Kitts, Shengjun Pan, Junwei Pan, and Aaron Flores. 2020. Bid shading in the brave new world of first-price auctions. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2453–2460.

[11] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. *Advances in Neural Information Processing Systems* 32 (2019).

[12] Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. 2020. Optimal no-regret learning in repeated first-price auctions. *arXiv preprint arXiv:2003.09795* (2020).

[13] Chong Jiang. 2015. *Online advertisements and multi-armed bandits*. University of Illinois at Urbana-Champaign.

[14] Niklas Karlsson and Qian Sang. 2021. Adaptive bid shading optimization of first-price ad inventory. In *2021 American Control Conference (ACC)*. IEEE, 4983–4990.

[15] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2012. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*. PMLR, 592–600.

[16] Xu Li, Michelle Ma Zhang, Zhenya Wang, and Youjun Tong. 2022. Arbitrary distribution modeling with censorship in real-Time bidding Advertising. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3250–3258.

[17] Weichao Mao, Zhenzhe Zheng, Fan Wu, and Guihai Chen. 2018. Online pricing for revenue maximization with unknown time discounting valuations. In *IJCAI*. 440–446.

[18] Shengjun Pan, Brendan Kitts, Tian Zhou, Hao He, Bharatbhushan Shetty, Aaron Flores, Djordje Gligorijevic, Junwei Pan, Tingyu Mao, San Gultekin, et al. 2020. Bid shading by win-rate estimation and surplus maximization. *arXiv preprint arXiv:2009.09259* (2020).

[19] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, and Yong Yu. 2019. Deep landscape forecasting for real-time bidding advertising. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 363–372.

[20] Jason Rhuggenaath, Alp Akcay, Yingqian Zhang, and Uzay Kaymak. 2022. Setting reserve prices in second-price auctions with unobserved bids. *INFORMS Journal on Computing* 34, 6 (2022), 2950–2967.

[21] Amin Sayedi. 2018. Real-time bidding in online display advertising. *Marketing Science* 37, 4 (2018), 553–568.

[22] Eric M Schwartz, Eric T Bradlow, and Peter S Fader. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* 36, 4 (2017), 500–522.

[23] Aleksandrs Slivkins et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* 12, 1-2 (2019), 1–286.

[24] Leonard J Tashman. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* 16, 4 (2000), 437–450.

[25] Tuomo Tilli and Leonardo Espinosa-Leal. 2021. Multi-armed bandits for bid shading in first-price real-time bidding auctions. *Journal of Intelligent & Fuzzy Systems* Preprint (2021), 1–15.

[26] Wush Wu, Mi-Yen Yeh, and Ming-Syan Chen. 2018. Deep censored learning of the winning price in the real time bidding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2526–2535.

[27] Wush Chi-Hsuan Wu, Mi-Yen Yeh, and Ming-Syan Chen. 2015. Predicting winning price in real time bidding with censored data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1305–1314.

[28] Congde Yuan, Mengzhuo Guo, Chaoneng Xiang, Shuangyang Wang, Guoqing Song, and Qingpeng Zhang. 2022. An actor-critic reinforcement learning model for optimal bidding in online display advertising. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3604–3613.

[29] Wei Zhang, Brendan Kitts, Yanjun Han, Zhengyuan Zhou, Tingyu Mao, Hao He, Shengjun Pan, Aaron Flores, San Gultekin, and Tsachy Weissman. 2021. MEOW: A space-efficient nonparametric bid shading algorithm. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3928–3936.

[30] Tian Zhou, Hao He, Shengjun Pan, Niklas Karlsson, Bharatbhushan Shetty, Brendan Kitts, Djordje Gligorijevic, San Gultekin, Tingyu Mao, Junwei Pan, et al. 2021. An efficient deep distribution network for bid shading in first-price auctions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3996–4004.