BANDITS MEET MECHANISM DESIGN TO COMBAT CLICKBAIT IN ONLINE RECOMMENDATION

Thomas Kleine Buening[1], **Aadirupa Saha**[2*], **Christos Dimitrakakis**[3], **Haifeng Xu**[4]
[1]The Alan Turing Institute, [2]TTIC, [3]University of Neuchatel, [4]University of Chicago

ABSTRACT

We study a strategic variant of the multi-armed bandit problem, which we coin the *strategic click-bandit*. This model is motivated by applications in online recommendation where the choice of recommended items depends on both the click-through rates and the post-click rewards. Like in classical bandits, rewards follow a fixed unknown distribution. However, we assume that the click-rate of each arm is chosen strategically by the arm (e.g., a host on Airbnb) in order to maximize the number of times it gets clicked. The algorithm designer does not know the post-click rewards nor the arms' actions (i.e., strategically chosen click-rates) in advance, and must learn both values over time. To solve this problem, we design an incentive-aware learning algorithm, UCB-S, which achieves two goals simultaneously: (a) incentivizing desirable arm behavior under uncertainty; (b) minimizing regret by learning unknown parameters. We approximately characterize all Nash equilibria of the arms under UCB-S and show a $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret bound uniformly in *every* equilibrium. We also show that incentive-unaware algorithms generally fail to achieve low regret in the strategic click-bandit. Finally, we support our theoretical results by simulations of strategic arm behavior which confirm the effectiveness and robustness of our proposed incentive design.

1 INTRODUCTION

Recommendation platforms act as intermediaries between *vendors* and *users* so as to recommend *items* from the former to the latter. On Amazon, vendors sell physical items, while on Youtube the recommended items are videos. The recommendation problem is how to select one or more items to present to each user so that they are most likely to click on at least one of them.

However, vendor-chosen *item descriptions* are an essential aspect of the problem that is often ignored. These invite vendors to exaggerate their true value in the descriptions in order to increase their Click-Through-Rates (CTRs). As a consequence, even though online learning algorithms can generally identify relevant items, the existence of unrepresentative or exaggerated item descriptions remains a challenge (Yue et al., 2010; Hofmann et al., 2012). These include thumbnails or headlines that do not truly reflect the underlying item (see Figure 1)—a well-known internet phenomenon called the *clickbait* (Wang et al., 2021). While moderately increasing user click-rates through attractive descriptions is often encouraged since it helps to increase the overall user activity, clickbait can be harmful to a platform as it leads to bad recommendation outcomes and damage to the platform's reputation which may exceed the value of any additional clicks. A key reason for such dishonest or exaggerated item deceptions is the *strategic behavior* of vendors driven by their incentive to increase their item's exposure and click probability. Thus naturally, vendors are better off carefully choosing descriptions so as to increase click-rates, which leads to phenomena such as clickbait.[1]

To address this issue, we take an approach that marries *mechanism design* without payments with *online learning*, which are two celebrated research areas, however, mostly studied as separate streams. Since clickbait is fundamentally driven by vendor incentives, we believe that the novel design of online learning policies *that can carefully align vendor incentives with the platform's overall objective* may help to resolve this issue from its root.

---

[*] Author is currently with Apple ML Research.

[1] This is possible because most platforms rely on vendors to provide descriptions about their items. For instance, the images of restaurants on Yelp, rentals on Airbnb, hotels on Expedia, title and thumbnails of Youtube videos, and descriptions of products on Amazon are all provided by the vendors.

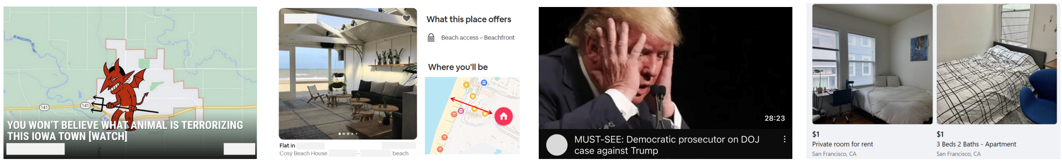

YOU WON'T BELIEVE WHAT ANIMAL IS TERRORIZING THIS IOWA TOWN [WATCH]
[redacted]
What this place offers
Beach access – Beachfront
Where you'll be
Flat in [redacted]
Cosy Beach House [redacted] - [redacted] beach
28:23
MUST-SEE: Democratic prosecutor on DOJ case against Trump
$1
Private room for rent
San Francisco, CA
$1
3 Beds 2 Baths - Apartment
San Francisco, CA

Figure 1: Examples of unrepresentative or clickbait headlines and thumbnails on Bing News, Airbnb, Youtube, and Facebook Marketplace (identifying information partly redacted).

To incorporate vendor-chosen item descriptions in this setting, we propose and study a natural strategic variant of the classical Multi-Armed Bandit (MAB) problem, which we call the *strategic click-bandit* in order to emphasize the strategic role that clicks and CTRs play in our setup.[2] Concretely, in strategic click-bandits, each arm i is characterized by (a) a reward distribution with mean μ_i, inherent to the arm; and (b) a click probability $s_i \in [0, 1]$, chosen freely by the arm at the beginning. Since the learner (i.e., the recommendation system) knows neither of these values in advance, it must learn them through interaction. The learner's objective is represented through a general utility function $u(s_i, \mu_i)$ that depends on both click-rate and post-click rewards.

We highlight two fundamental differences between strategic click-bandits and standard MABs. First, each arm in the strategic click-bandit is a *self-interested agent* whose objective is to maximize the number of times it gets clicked. This captures the strategic behavior of many vendors in online recommendations, especially those who are rewarded based on user clicks (e.g., Youtube (2023)). Second, s_i is a freely chosen *action* by arm i, rather than a fixed parameter of arm i. We believe these modeling adjustments more realistically capture vendor behaviors in real applications. They also lead to intriguing mechanism design questions since the bandit algorithm not only needs to learn the unknown parameters, but also has to carefully align incentives to avoid undesired arm behavior. In summary, our contributions are:

1. We introduce the strategic click-bandit problem, which involves strategic arms manipulating click-rates so as to maximize their own utility, and show that *incentive-unaware* algorithms generally fail to achieve low regret in the strategic click-bandit (Section 3, Proposition 4.1).
2. We design an *incentive-aware* learning algorithm, UCB-S, that combines mechanism design and online learning techniques and effectively incentivizes desirable arm strategies while minimizing regret by making credible and justified threats to arms under uncertainty (Section 5).
3. We characterize the set of Nash equilibria for the arms under the UCB-S mechanism and show that every arm i's strategy is $\widetilde{\mathcal{O}}\big(\max\big\{\Delta_i, \sqrt{K/T}\big\}\big)$ close to the desired strategy in equilibrium (Theorem 5.2). We then show that UCB-S achieves $\widetilde{\mathcal{O}}\big(\sqrt{KT}\big)$ strong strategic regret (Theorem 5.3) and complement this with an almost matching lower bound of $\Omega\big(\sqrt{KT}\big)$ for weak strategic regret (Theorem 5.5).
4. We simulate strategic arm behavior through repeated interaction and gradient ascent and empirically demonstrate the effectiveness of the proposed UCB-S mechanism (Section 6).

2 RELATED WORK

The MAB problem is a well-studied online learning framework, which can be used to model decision-making under uncertainty (Lai et al., 1985; Auer, 2002). Since it inherently involves sequential actions and the exploration-exploitation trade-off, the MAB framework has been applied to online recommendations (Li et al., 2010; Zong et al., 2016; Wang et al., 2017) as well as a myriad of other domains (Bouneffouf et al., 2020). While there is much work studying strategic machine learning (e.g., Hardt et al., 2016; Freeman et al., 2020; Zhang and Conitzer, 2021), we here wish to highlight related work that connects online learning (and specifically the MAB formalism) to mechanism design (Nisan and Ronen, 1999). Additional related work is discussed in Appendix H.

To the best of our knowledge, Braverman et al. (2019) are the first to study a strategic variant of the MAB problem. In their model, when an arm is pulled, it receives a privately observed reward ν and chooses to pass on a portion x of it to the principal, keeping $\nu - x$ for itself. The goal of

---

[2] We use the terms click-through-rate, click-rate, and click probability interchangeably.

Model 1: The Strategic Click-Bandit Problem

```
Learner commits to algorithm $M$, which is shared with all arms
Arms choose strategies $(s_1, \dots, s_K) \in [0,1]^K$ (unknown to $M$)
for $t = 1, \dots, T$ do
    Algorithm $M$ selects arm $i_t \in [K]$
    Arm $i_t$ is clicked with probability $s_{i_t}$, i.e., $c_{t,i_t} \sim \mathrm{Bern}(s_{i_t})$
    if $i_t$ was clicked ($c_{t,i_t} = 1$) then
        Arm $i_t$ receives utility $1$ from the click
        $M$ observes post-click reward $r_{t,i_t}$ drawn from a distribution with mean $\mu_{i_t}$
```

the principal is then to incentivize arms to share as much reward with the principal as possible. In contrast to our work, the principal must not learn the underlying reward distribution or the arm strategies, but instead design an auction among arms based on the shared rewards. Feng et al. (2020) and Dong et al. (2022) study the robustness of bandit algorithms to strategic reward manipulations. However, neither work attempts to align incentives by designing mechanisms, but instead assume a limited manipulation budget. Shin et al. (2022) study MABs with strategic replication in which agents can submit several arms with replicas to the platform. They design an algorithm, which separately explores the arms submitted by each agent and in doing so discourages agents from creating additional arms and replicas. Another line of work studies auction-design in MAB formalisms, often motivated by applications in ad auctions (Babaioff et al., 2009; Devanur and Kakade, 2009; Babaioff et al., 2015). In these models, in every round the auctioneer selects one advertiser's item, which is subsequently clicked or not, and the goal of the auctioneer is to incentivize advertisers to truthfully bid their value-per-click by constructing selection and payment rules.

To the best of our knowledge, our work is the first to study the situation where the arms' strategies (as well as other parameters) are initially unobserved, and must be learned from interaction while simultaneously incentivizing arms under uncertainty without payments. As a result, while other work is usually able to precisely incentivize certain arm strategies, our mechanism design and characterization of the Nash equilibria are *approximate*.

3 THE STRATEGIC CLICK-BANDIT PROBLEM

We consider a natural strategic variant of the classical MAB, motivated by applications in online recommendation. Unlike classical MABs, strategic click-bandits feature decentralized interactions with the learner and multiple self-interested arms.

Let $[K] := \{1, \dots, K\}$ denote the set of arms, each being viewed as a strategic *agent*. The strategic click-bandit proceeds in two phases. In the first phase, the learner commits to an online learning policy M, upon which each arm i chooses a description, which results in a corresponding click-rate $s_i \in [0,1]$. The second phase proceeds in rounds. At each round t: (1) the algorithm M pulls/recommends an arm i_t based on observed past data; (2) arm i_t is clicked with probability s_{i_t}; (3) if i_t is clicked, arm i_t receives utility 1 (whereas all other arms i receive utility 0) and the learner observes a post-click reward $r_{t,i_t} \in [0,1]$ drawn from i_t's reward distribution with mean $\mu_{i_t} \in [0,1]$. If i_t is *not* clicked, all arms receive 0 utility and the learner does not observe any post-click rewards. The post-click mean μ_i is fixed for each arm i and captures the *true value* of the arm. From the learner's perspective, *both* s_i and μ_i of each arm are unknown but can be learned from online bandit feedback, that is, whether the recommended arm is clicked and, if so, what its realized reward is. In the following, we will also refer to the online learning policy M as a *mechanism* to emphasize its dual role in learning and incentive design. We summarize the interaction in Model 1.

3.1 LEARNER'S UTILITY

The learner's utility of selecting an arm i with CTR s_i and post-click value μ_i is denoted $u(s_i, \mu_i)$. One example of this utility function is $u(s, \mu) = s\mu$. In this case, the learner monotonically prefers large s and does not care about how much the click-rate s differs from the post-click value μ. However, we believe that the learner (e.g., a platform like Youtube or Airbnb) usually values consistency between the click-rates and the post-click values of arms. This could be captured by a penalty term for how much s_i differs from μ_i; for instance, a natural choice is $u(s, \mu) = s\mu - \lambda(s - \mu)^2$ for some weight $\lambda > 0$. Such *non-monotonicity* of the learner's utility $u(s_i, \mu_i)$ in s_i *versus* arm i's monotonic preference of larger click-rates forms the fundamental tension in the strategic click-bandit model and is also the reason that mechanism design is needed. We keep the above utility functions in mind as running examples, but derive our results for a much more general class of functions satisfying the following mild regularity assumptions:

(A1) $u\colon [0, 1] \times [0, 1] \to \mathbb{R}$ is L-Lipschitz w.r.t. the ℓ_1-norm.

(A2) $u^*(\mu) := \max_{s\in[0,1]} u(s, \mu)$ is monotonically increasing.

(A3) $s^*(\mu) := \operatorname{argmax}_{s\in[0,1]} u(s, \mu)$ is H-Lipschitz and is bounded away from zero.

Assumption (A1) bounds the loss of selecting a suboptimal arm. (A2) states that, in the (idealized) situation when the arms choose click-rates so as to maximize the learner's utility u, then arms with larger post-click rewards μ are always preferred. (A3) then ensures that from the perspective of the learner most desired strategy $s^*(\mu)$ does not change abruptly w.r.t. μ and the learner wishes to incentivize non-zero click-rates. In what follows, the function $s^*(\mu)$ will play a central role as it describes the arm strategy that maximizes the learner's utility. For instance, in the case of $u(s, \mu) = s\mu - \lambda(s - \mu)^2$ it is given by $s^*(\mu) = (1 + \frac{1}{2\lambda})\mu$. As such, the learner will typically try to incentivize an arm with post-click reward μ_i to choose strategy $s^*(\mu_i)$.

3.2 ARMS' UTILITY AND NASH EQUILIBRIA AMONG ARMS

The mean post-click reward μ_i of each arm i is fixed, whereas arm i can freely choose the CTR s_i. In the strategic click-bandit, the objective of each arm i is to maximize the number of times it gets clicked $\sum_{t=1}^{T} \mathbb{1}_{\{i_t=i\}}\, c_{t,i}$, which captures the objectives of vendors on internet platforms for whom user traffic typically proportionally converts to revenue.[3] We now introduce the solution concept for the game among arms defined by a mechanism M and post-click rewards $\mu_1, \dots, \mu_K$, often referred to as an *equilibrium*. Let s_{-i} denote the $K - 1$ strategies of all arms except i. Each arm i chooses s_i to maximize their *expected* number of clicks $v_i(M, s_i, s_{-i})$, which is a function of the mechanism M, their own action s_i as well as all other arms' actions s_{-i}. Concretely,

$$v_i(M, s_i, s_{-i}) := \mathbb{E}_M\left[\sum_{t=1}^{T} \mathbb{1}_{\{i_t=i\}}\, c_{t,i}\right] \tag{1}$$

where the expectation is taken over the mechanism's decisions and the environment's randomness. We generally write $\boldsymbol{s} := (s_1, \dots, s_K)$ to summarize a strategy profile of the arms. Let Σ denote the set of probability measures over $[0, 1]$. Given a *mixed* strategy profile $\boldsymbol{\sigma} = (\sigma_i, \sigma_{-i}) \in \Sigma^K$, i.e., a distribution over $[0, 1]^K$, arm i's utility is then defined as $v_i(M, \sigma_i, \sigma_{-i}) := \mathbb{E}_{\boldsymbol{s}\sim\boldsymbol{\sigma}}[v_i(M, s_i, s_{-i})]$.

Definition 3.1 (Nash Equilibrium). We say that $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_K) \in \Sigma^K$ is a Nash equilibrium (NE) under mechanism M if $v_i(M, \sigma_i, \sigma_{-i}) \geq v_i(M, \sigma_i', \sigma_{-i})$ for all $i \in [K]$ and strategies $\sigma_i' \in \Sigma$.

In other words, $\boldsymbol{\sigma}$ is in NE if no arm can increase its utility by *unilaterally* deviating to some other strategy. If some NE $\boldsymbol{\sigma} \in \Sigma^K$ has weight one on a pure strategy profile $\boldsymbol{s} \in [0, 1]^K$, this equilibrium is said to be in pure-strategies. Let $\mathrm{NE}(M) := \{\boldsymbol{\sigma} \in \Sigma^K : \boldsymbol{\sigma} \text{ is a NE under } M\}$ denote the set of all (possibly mixed) NE under mechanism M. Following conventions in standard economic analysis, we assume that the arms will form a NE in $\mathrm{NE}(M)$ in response to an algorithm M.[4]

Remark 3.1 (Existence of Nash Equilibrium). *In general, the arms' utility functions $v_i(M, s_i, s_{-i})$ may be discontinuous in the arms' strategies due to their intricate dependence on the learning algorithm M. It is well-known that in games with discontinuous utilities, a NE may not exist (Reny, 1999). However, for all subsequently considered algorithms we will prove the existence of a NE by either explicitly describing the equilibrium or implicitly proving its existence.*

---

[3] More generally, different arms i may have a different value-per-click ν_i that could as well depend on μ_i so that $v_i(M, s_i, s_{-i}) = \mathbb{E}_M[\sum_t \mathbb{1}_{\{i_t=i\}}\, c_{t,i}\, \nu_i]$. This can easily be accommodated for by our model and our results readily extend to this case since each arm's goal still boils down to maximizing the number of clicks.

[4] For instance, a sufficient condition for the arms to find a NE is their knowledge about how far away they are from the best arm, i.e., their optimality gap in post-click rewards $\Delta_i := \max_{j\in[K]} \mu_j - \mu_i$.

3.3 STRATEGIC REGRET

The learner's goal is to maximize $\sum_{t=1}^{T} u(s_{i_t}, \mu_{i_t})$ which naturally depends on the arm strategies $s_1, \dots, s_K$. For given post-click values $\mu_1, \dots, \mu_K$, the maximal utility $u(s^*, \mu^*)$ is then achieved for $\mu^* := \max_{i \in [K]} \mu_i$ and $s^* := s^*(\mu^*)$, that is, $u(s^*, \mu^*) = \max_{i \in [K]} \max_{s \in [0,1]} u(s, \mu_i)$. With $u(s^*, \mu^*)$ as a benchmark, we can define the *strategic regret* of a mechanism M under a pure-strategy equilibrium $\boldsymbol{s} \in \mathrm{NE}(M)$ as

$$R_T(M, \boldsymbol{s}) := \mathbb{E}\left[\sum_{t=1}^{T} u(s^*, \mu^*) - u(s_{i_t}, \mu_{i_t})\right]. \tag{2}$$

For some mixed-strategy equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(M)$, we then accordingly define strategic regret as $R_T(M, \boldsymbol{\sigma}) := \mathbb{E}_{\boldsymbol{s} \sim \boldsymbol{\sigma}}[R_T(M, \boldsymbol{s})]$. In general, there may exist several Nash equilibria for the arms under a given mechanism M. We can then consider the *strong strategic regret* of M given by the regret under the worst-case equilibrium:

$$R_T^+(M) := \max_{\boldsymbol{\sigma} \in \mathrm{NE}(M)} R_T(M, \boldsymbol{\sigma}),$$

or the *weak strategic regret* given by the regret under the most favorable equilibrium:

$$R_T^-(M) := \min_{\boldsymbol{\sigma} \in \mathrm{NE}(M)} R_T(M, \boldsymbol{\sigma}),$$

where $R_T^-(M) \le R_T^+(M)$. The regret upper bound of our proposed algorithm, UCB-S, holds under any equilibrium in NE(UCB-S), thereby bounding *strong strategic regret* (Theorem 5.3). On the other hand, the proven lower bounds (Proposition 4.1 and Theorem 5.5) hold for *weak strategic regret* and thus also apply to its strong counterpart.

4 LIMITATIONS OF INCENTIVE-UNAWARE ALGORITHMS

We start our analysis of the strategic click-bandit problem by showing that simply finding the arm with the largest post-click reward, $\mathrm{argmax}_i\, \mu_i$, or largest utility, $\mathrm{argmax}_i\, u(s_i, \mu_i)$, is insufficient to achieve $o(T)$ *weak* strategic regret. In fact, we find that even with oracle knowledge of $\mu_1, \dots, \mu_K$ and $s_1, \dots, s_K$, an algorithm may suffer linear weak strategic regret if it fails to account for the arms' strategic nature. For such incentive-*unaware* oracle algorithms, we show a $\Omega(T)$ lower bound for weak strategic regret on any non-trivial problem instance.

Recall that $\mu^* := \max_{i \in [K]} \mu_i$ and $s^* := s^*(\mu^*)$ and suppose that the arm $i^* = \mathrm{argmax}_{i \in [K]}\, \mu_i$ with maximal post-click rewards is unique. Our negative results rely on the following problem-dependent gaps in terms of utility:

$$\beta := u(s^*, \mu^*) - u(1, \mu^*) \quad \text{and} \quad \eta := u(s^*, \mu^*) - \max_{i \in [K] \setminus \{i^*\}} u^*(\mu_i).$$

Here, β denotes the cost of the optimal arm i^* deviating from the desired strategy $s^* = s^*(\mu^*)$ by playing $s_{i^*} = 1$. The quantity η denotes the gap between the maximally achievable utility $u(s^*, \mu^*)$ and the utility of the second best arm.

Proposition 4.1. *Let μ-Oracle be the algorithm with oracle knowledge of $\mu_1, \dots, \mu_K$ that plays $i_t = \mathrm{argmax}_{i \in [K]}\, \mu_i$ in every round t, whereas (s, μ)-Oracle is the algorithm with oracle knowledge of $\mu_1, \dots, \mu_K$ and $s_1, \dots, s_K$ that always plays $i_t = \mathrm{argmax}_{i \in [K]} u(s_i, \mu_i)$ with ties broken in favor of the larger μ. We then have*

(i) Under every equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\mu\text{-}Oracle)$, the μ-Oracle suffers regret $\Omega\big(\beta T\big)$, i.e.,

$$R_T^-(\mu\text{-}Oracle) = \Omega\big(\beta T\big).$$

(ii) Under every $\boldsymbol{\sigma} \in \mathrm{NE}((s, \mu)\text{-}Oracle)$, the (s, μ)-Oracle suffers regret $\Omega\big(\min\{\beta, \eta\} T\big)$, i.e.,

$$R_T^-((s, \mu)\text{-}Oracle) = \Omega\big(\min\{\beta, \eta\} T\big).$$

Mechanism 1: UCB with Screening (UCB-S)

```
1  initialize: $A_0 = [K]$
2  for $t = 1, \dots, T$ do
3      if $A_{t-1} \neq \emptyset$ then
4          Select $i_t \in \operatorname{argmax}_{i \in A_{t-1}} \overline{\mu}_i^{t-1}$
5      else
6          Select $i_t$ uniformly at random from $[K]$
7      Arm $i_t$ is clicked with probability $s_{i_t}$, i.e., $c_{t,i_t} \sim \mathrm{Bern}(s_{i_t})$
8      if $i_t$ was clicked ($c_{t,i_t} = 1$) then
9          Observe post-click reward $r_{t,i_t}$
10     if $\overline{s}_{i_t}^t < \min_{\mu \in [\underline{\mu}_{i_t}^t, \overline{\mu}_{i_t}^t]} s^*(\mu)$ or $\underline{s}_{i_t}^t > \max_{\mu \in [\underline{\mu}_{i_t}^t, \overline{\mu}_{i_t}^t]} s^*(\mu)$ then
11         Ignore arm $i_t$ in future rounds: $A_t \leftarrow A_{t-1} \setminus \{i_t\}$
```

Proof Sketch. (i): We show that $s = 1$ is a strictly dominant strategy for arm i^* under the μ-Oracle. This implies that arm i^* plays $s_{i^*} = 1$ with probability one in every NE under the μ-Oracle. The claimed lower bound then follows from bounding the instantaneous regret per round from below by β. *(ii)*: Let $j^* \in \operatorname{argmax}_{i \neq i^*} \mu_i$. It can be seen that in any NE, arm i^* will play the largest $s \in [0, 1]$ such that $u(s, \mu_{i^*}) \geq u(s_{j^*}, \mu_{j^*})$. We then show that either $s_{i^*} = 1$ or $u(s_{i^*}, \mu_{i^*}) = u(s^*(\mu_{j^*}), \mu_{j^*})$. Once again this allows us to lower bound the regret per round by $\min\{\beta, \eta\}$. $\square$

As a concrete example of the failure of the μ-Oracle and the (s, μ)-Oracle, let us consider the running example of $u(s, \mu) = s\mu - \lambda(s - \mu)^2$. In this case, letting $\lambda = 5$ and $\mu_{i^*} = 0.8$ and $\mu_i \leq 0.7$ for $i \neq i^*$, we get $\beta \geq 0.1$ and $\eta \geq 0.1$ so that both oracles suffer $\Omega(T)$ regret in every equilibrium.

5 NO-REGRET INCENTIVE-AWARE LEARNING: UCB-S

The results of Proposition 4.1 suggest that any incentive-unaware learning algorithm that is oblivious to the strategic nature of the arms will generally fail to achieve low regret. In particular, "unconditional" selection of any arm will likely result in undesirable equilibria among arms. For these reasons, we deploy a conceptually simple screening idea, which threatens arms with elimination when deviating from the desired strategies.

Let denote $n_t(i)$ be the number of times up to (and including) round t that arm i was selected by the learner, and let $m_t(i)$ denote the number of times post-click rewards were observed for arm i up to (and including) round t. Let $\widehat{s}_i^t$ be the average observed click-rate and $\widehat{\mu}_i^t$ the average observed post-click reward for arm i. We then define the pessimistic and optimistic estimates of s_i and μ_i as

$$\underline{s}_i^t = \widehat{s}_i^t - \sqrt{2\log(T)/n_t(i)}, \qquad \overline{s}_i^t = \widehat{s}_i^t + \sqrt{2\log(T)/n_t(i)},$$
$$\underline{\mu}_i^t = \widehat{\mu}_i^t - \sqrt{2\log(T)/m_t(i)}, \qquad \overline{\mu}_i^t = \widehat{\mu}_i^t + \sqrt{2\log(T)/m_t(i)}.$$

where $\underline{s}_i^t = -\infty$ and $\overline{s}_i^t = +\infty$ for $n_t(i) = 0$ as well as $\underline{\mu}_i^t = -\infty$ and $\overline{\mu}_i^t = +\infty$ for $m_t(i) = 0$.

In every round, UCB-S (Mechanism 1) selects arms optimistically according to their post-click rewards and subsequently observes if the arm is clicked, i.e., c_{t,i_t}, and, if so, a post-click reward r_{t,i_t}. However, if an arm's click-rate s_i is detected to be different from the learner's desired arm strategy $s^*(\mu_i)$, the arm is eliminated forever, expressed by the screening rule in line 10:

$$\overline{s}_{i_t}^t < \min_{\mu \in [\underline{\mu}_{i_t}^t, \overline{\mu}_{i_t}^t]} s^*(\mu) \quad \text{or} \quad \underline{s}_{i_t}^t > \max_{\mu \in [\underline{\mu}_{i_t}^t, \overline{\mu}_{i_t}^t]} s^*(\mu).$$

The only exception is when all arms have been eliminated. Then, UCB-S plays them all uniformly for the remaining rounds. To ensure that the elimination of an arm is credible and justified with high probability, we leverage confidence bounds on s_i and μ_i. More precisely, if an arm is truthful and chooses $s_i = s^*(\mu_i)$, then with probability $1 - 1/T^2$ it will not be eliminated by the screening rule.

As a prelude to the analysis of the UCB-S mechanism, we begin by showing that there always exists a NE among the arms under UCB-S. As mentioned briefly in Section 3, the existence of a NE among the arms is not guaranteed under an arbitrary mechanism due to the arms' continuous strategy space and possibly discontinuous utility function.

Lemma 5.1. *For any post-click rewards $\mu_1, \dots, \mu_K$, there always exists a (possibly mixed) Nash equilibrium for the arms under the UCB-S mechanism.*

5.1 CHARACTERIZING THE NASH EQUILIBRIA UNDER UCB-S

We now approximately characterize all NE for the arms under the UCB-S mechanism. In order to prove a regret upper bound for UCB-S, it will be key to ensure that each arm i plays a strategy s_i which is sufficiently close to the desired strategy $s^*(\mu_i)$ (i.e., the strategy that maximizes the learner's utility). This is particularly important for arms i^* with maximal post-click rewards $\mu_{i^*} = \max_{i \in [K]} \mu_i$. If such arms i^* were to deviate substantially from $s^*(\mu_{i^*})$, e.g., by a constant amount, the learner would be forced to suffer constant regret even when selecting arms with maximal post-click rewards, making it impossible to achieve sublinear regret.

In the following, we show that under the UCB-S mechanism every NE is such that the strategies of arms with maximal post-click rewards deviate from the desired strategies by at most $\widetilde{\mathcal{O}}(\sqrt{K/T})$. We then also show that for suboptimal arms the difference between each arm i's strategy s_i and the desired strategy $s^*(\mu_i)$ is governed by their optimality gap in post-click rewards, given by $\Delta_i := \mu^* - \mu_i$. Recall that H denotes the Lipschitz constant of $s^*(\mu)$.

Theorem 5.2. *For all $\boldsymbol{s} \in \operatorname{supp}(\boldsymbol{\sigma})$ with $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$ and all $i \in [K]$:*

$$s_i = s^*(\mu_i) + \mathcal{O}\left(H \cdot \max\left\{\Delta_i, \sqrt{\frac{K \log(T)}{T}}\right\}\right).$$

In particular, for all arms $i^* \in [K]$ with $\Delta_{i^*} = 0$, i.e., maximal post-click rewards:

$$s_{i^*} = s^*(\mu_{i^*}) + \mathcal{O}\left(H \sqrt{\frac{K \log(T)}{T}}\right).$$

The derivation of Theorem 5.2 can be best understood by noting that the estimates of each arm's strategy roughly concentrate at a rate of $1/\sqrt{t}$. Then, depending on how often an arm expects to be selected by UCB-S, it can exploit our uncertainty about its strategy and safely increase its click-rates to match our confidence. Generally, optimal arms expect at least T/K allocations while preventing elimination, which can be seen to imply NE strategies that deviate by at most $\sqrt{K/T}$. On the other hand, suboptimal arms can expect roughly $\log(T)/\Delta_i^2$ allocations as long as they can prevent elimination and all other arms act rationally, which results in the linear dependence on Δ_i. Hence, interestingly UCB-S' selection policy directly impacts the truthfulness of the arms, as arms that are selected more frequently are forced to choose strategies closer to $s^*(\mu_i)$. We thus observe a trade-off between incentivizing *all* arms to be truthful and recommending only the best arms. The proof of Theorem 5.2 (Appendix C) then relies on the above observation and careful and repeated application of the best response property of the Nash equilibrium.

5.2 UPPER BOUND OF THE STRONG STRATEGIC REGRET OF UCB-S

With the approximate NE characterization from Theorem 5.2 at our disposal, we are ready to prove a regret upper bound for UCB-S. We show that the *strong strategic regret* of the UCB-S mechanism is upper bounded by $\widetilde{\mathcal{O}}(\sqrt{KT})$, that is, for any $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$ the regret guarantee holds.

Theorem 5.3. *Let $\Delta_i := \mu^* - \mu_i$ and let L and H denote the Lipschitz constants of $u(s, \mu)$ and $s^*(\mu)$, respectively. The strong strategic regret of UCB-S is bounded as*

$$R_T^+(\text{UCB-S}) = LH \cdot \mathcal{O}\left(\sqrt{KT\log(T)} + \sum_{i:\Delta_i > 0} \frac{\log(T)}{\Delta_i}\right). \tag{3}$$

In other words, the above regret bound is achieved under any equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$.

Proof Sketch. As suggested by the regret bound there are two sources of regret. Broadly speaking, the first term on the right hand side of (3) corresponds to the regret UCB-S suffers due to arms with maximal post-click rewards (i.e., $\Delta_i = 0$) deviating from the utility-maximizing strategy $s^*(\mu^*)$. For such arms Theorem 5.2 bounded the deviation by a term of order $\sqrt{K/T}$, thereby leading to at most order $\sqrt{KT}$ regret. The second term in (3) corresponds to the regret suffered from playing arms with suboptimal post-click rewards, i.e., $\Delta_i > 0$. Using a typical UCB argument, the Lipschitzness of $u(s, \mu)$ and $s^*(\mu)$, and again Theorem 5.2 applied to $|s^*(\mu^*) - s_i| \leq |s^*(\mu^*) - s^*(\mu_i)| + \mathcal{O}(H\Delta_i) \leq H\Delta_i + \mathcal{O}(H\Delta_i)$ we obtain the claimed upper bound. □

Similarly to classical MABs we can state a regret bound independent of the instance-dependent quantities Δ_i and translate Theorem 5.3 into a minimax-type guarantee.

Corollary 5.4. *The strong strategic regret of UCB-S is bounded as*

$$R_T^+(\text{UCB-S}) = \mathcal{O}\left(LH\sqrt{KT\log(T)}\right).$$

In other words, the above regret bound is achieved under any equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$.

Theorem 5.3 nicely shows that the additional cost of the incentive design and the strategic behavior of the arms is of order $\sqrt{KT}$ which primarily stems from arms with maximal post-click rewards deviating by roughly $\sqrt{K/T}$ from the desired strategy (see Theorem 5.2). The dishonesty of suboptimal arms does not notably contribute to the regret and is contained in the $\log(T)/\Delta_i$ expressions as we can bound the number of times suboptimal arms are played sufficiently well. As a result, the total cost of incentive design and strategic behavior matches the minimax learning complexity of MABs so that we obtain an overall $\widetilde{\mathcal{O}}(\sqrt{KT})$ strategic regret bound under every equilibrium.

5.3 LOWER BOUND FOR WEAK STRATEGIC REGRET

Complementing our regret analysis, we prove a lower bound on *weak strategic regret* in the strategic click-bandit. By definition, weak strategic regret lower bounds its strong counterpart, i.e., $R_T^-(M) \leq R_T^+(M)$, so that the shown lower bound directly applies to strong strategic regret as well, which implies that UCB-S is near-optimal.

Theorem 5.5. *Let* M *be any mechanism with* $\mathrm{NE}(M) \neq \emptyset$*. There exists a utility function* u *satisfying (A1)-(A3) and post-click rewards* $\mu_1, \dots, \mu_K$ *such that for all Nash equilibria* $\boldsymbol{\sigma} \in \mathrm{NE}(M)$*:*

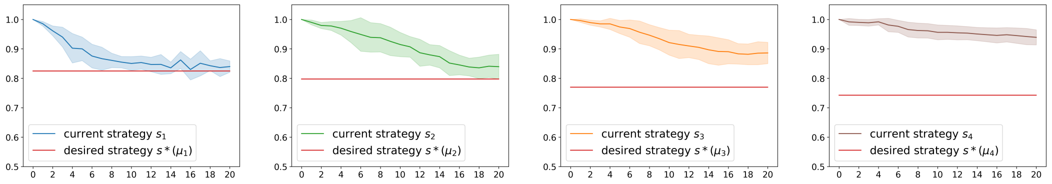
$$R_T(M, \boldsymbol{\sigma}) = \Omega\big(\sqrt{KT}\big).$$

In other words, $R_T^-(M) = \Omega\big(\sqrt{KT}\big)$.

Proof Sketch. Consider the utility function $u(s, \mu) = s\mu$. Intuitively, for any low regret mechanism M the NE for the arms will be in $(s_1, \dots, s_K) = (1, \dots, 1)$ as these strategies maximize the learner's utility u and are to the advantage of the arms. In this case, the learning problem reduces to a classical MAB and we inherit the well-known minimax $\sqrt{KT}$ lower bound. However, it is not directly clear that there exists no better mechanism that would, e.g., incentivize arm strategies $(s_1, \dots, s_{i^*}, \dots, s_K) = (0, \dots, 1, \dots, 0)$ under which $i^* = \operatorname{argmax}_i \mu_i$ becomes easier to distinguish from $i \neq i^*$. For this reason, we argue via the arms' utilities and lower bound the minimal utility a suboptimal arm must receive in any NE. This directly implies a lower bound on the number of times we must play any suboptimal arm in equilibrium, which yields the claimed result. □

6 SIMULATING STRATEGIC ARM BEHAVIOR VIA REPEATED INTERACTION

Goal of the experiments is to analyze the effect of the proposed incentive-aware learning algorithm UCB-S on strategically responding arms. Strategic arm behavior is here modeled through decentralized gradient ascent and repeated interaction with the mechanism. Contrary to the assumption of arms playing in NE, arms follow a simple gradient ascent strategy to adapt to the mechanism, which serves as a realistic and natural model of strategic behavior. This requires no prior knowledge from the point of view of the arms and all learning is performed through sequential interaction with the mechanism. For this reason, the final strategies in our experiments may not necessarily be in NE. Despite this, we want to see whether the mechanism is still able to incentivize arms to behave in the desired manner which will also provide insight into the robustness of the proposed incentive design.



(a) Optimal arm with mean $\mu_1 = 0.75$.

(b) Suboptimal arm with mean $\mu_2 = 0.725$.

(c) Suboptimal arm with mean $\mu_3 = 0.7$.

(d) Suboptimal arm with mean $\mu_4 = 0.675$.

Figure 2: The strategic behavior of $K = 4$ arms when each arm uses gradient ascent to maximize their utility v_i in response to the UCB-S mechanism. In red, the desired strategy $s^*(\mu_i)$ for each arm i, respectively. As suggested by Theorem 5.2, the truthfulness, i.e., distance to $s^*(\mu_i)$, of a suboptimal arm i is governed by the arm's optimality gap Δ_i. We see this confirmed as the distance $s_i - s^*(\mu_i)$ increases as Δ_i increases. In accordance with our theoretical results, the optimal arm 1 has the largest incentive to play close to the desired strategy (as it loses the most when eliminated).

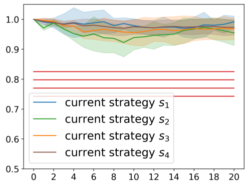


Figure 3: Strategic arm behavior when interacting with incentive-*unaware* standard UCB.

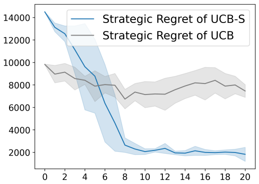


Figure 4: Strategic regret of UCB-S and standard UCB as arms adapt their strategies.

Experimental Setup. We consider the earlier introduced utility function defined as $u(s, \mu) = s\mu - \lambda(s - \mu)^2$ such that the desired (learner's utility-maximizing) strategy given μ is $s^*(\mu) = (1 + \frac{1}{2\lambda})\mu$. We let $\lambda = 5$. To model the strategic behavior of arms in response to UCB-S, we let the strategic arms interact with the mechanism over the course of 20 epochs (x-axis) and model each arm's strategic behavior via gradient ascent w.r.t. its utility v_i. More precisely, after every epoch (i.e., interaction over $T = 50$k rounds), each arm performs an approximated gradient step with respect to its utility v_i. We initialized the arm strategies to $s_i = 1$, however, our experiments show that other initialization, such as $s_i = 0$ or $s_i = 0.5$, yield similar results. All results are averaged over 10 complete runs and the standard deviation shown in shaded color.

Results. The conducted simulations show that under natural greedy behavior as modeled by gradient ascent, the incentive design of UCB-S is still effective and desirable arm strategies incentivized (Figure 2). Most notably, the optimal arm (having the largest incentive to be truthful) converges to a strategy close to the desired strategy $s^*(\mu_1)$. The suboptimal arms do not converge to a strategy close to the desired strategy and we observe that the distance to $s^*(\mu_i)$ depends on the optimality gap Δ_i, which mirrors our theoretical results (Theorem 5.2). In addition, Figure 4 shows that as the arms interact with UCB-S and adapt their strategies, the regret of UCB-S improves substantially. In contrast, incentive-unaware algorithms like UCB fail to incentivize desirable strategies (all arm strategies remain close to 1, see Figure 3) and UCB accordingly suffers large regret (Figure 4) throughout all epochs. The observation that UCB-S initially suffer larger regret than UCB can be explained by the elimination rule causing UCB-S to select arms uniformly at random when arms are notably untruthful. This threat of elimination, however, incentivizes the arms to adapt their strategies in the next epoch and eventually leads to smaller regret for UCB-S.

7 DISCUSSION

We study the strategic click-bandit problem in which each arm is associated with a click-rate, chosen strategically by the arms, and an immutable post-click reward. We show the necessity of incentive design in this model and design an incentive-aware online learning algorithm that incentivizes desirable arm strategies under uncertainty. As the learner has no prior knowledge of the arm strategies and the post-click rewards, the mechanism design is approximate and leaves room for arms to exploit the learner's uncertainty. This leads to an interesting regret bound which makes the intuition precise that arms can exploit the learner's uncertainty about their strategies. In our simulations we then observe that our incentive design is robust and still effective under natural greedy arm behavior and that the design of incentive-aware learning algorithms is necessary to achieve low regret under strategic arm behavior. Some interesting open questions which we leave for future work include whether the proposed incentive design remains effective under adaptive arm strategies and whether we can construct a mechanism under which there exists a desirable NE in dominant strategies.

Acknowledgments. We thank the anonymous reviewers for their insightful and constructive comments. We also want to thank Boi Faltings for helpful discussions in the early stages of this work. Thomas Kleine Buening was partly supported by the Norwegian Research Council (Grant 302203). Haifeng Xu is supported by the AI2050 program at Schmidt Sciences (Grant G-24-66104), an NSF Award CCF-2303372, an Army Research Office Award W911NF-23-1-0030, and an Office of Naval Research Award N00014-23-1-2802.

REFERENCES

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88, 2009.

Moshe Babaioff, Robert D Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. *Journal of the ACM (JACM)*, 62(2):1–37, 2015.

Dirk Bergemann and Juuso Välimäki. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–274, 2019.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

Mark Braverman, Jieming Mao, Jon Schneider, and S Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pages 383–416. PMLR, 2019.

Nikhil R Devanur and Sham M Kakade. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 99–106, 2009.

Jing Dong, Ke Li, Shuai Li, and Baoxiang Wang. Combinatorial bandits under strategic manipulations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 219–229, 2022.

Zhe Feng, David Parkes, and Haifeng Xu. The intrinsic robustness of stochastic bandits to strategic manipulation. In *International Conference on Machine Learning*, pages 3092–3101. PMLR, 2020.

Rupert Freeman, David M Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible prediction with expert advice. *arXiv preprint arXiv:2002.08837*, 2020.

Guoju Gao, He Huang, Mingjun Xiao, Jie Wu, Yu-E Sun, and Sheng Zhang. Auction-based combinatorial multi-armed bandit mechanisms with strategic arms. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Nicola Gatti, Alessandro Lazaric, and Francesco Trovò. A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 605–622, 2012.

Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 233–246, 2013.

Irving L Glicksberg. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Katja Hofmann, Fritz Behr, and Filip Radlinski. On caption bias in interleaving experiments. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 115–124, 2012.

Jiri Hron, Karl Krauth, Michael I Jordan, Niki Kilbertus, and Sarah Dean. Modeling content creator incentives on algorithm-curated platforms. *arXiv preprint arXiv:2206.13102*, 2022.

Xinyan Hu, Meena Jagadeesan, Michael I Jordan, and Jacob Steinhard. Incentivizing high-quality content in online recommender systems. *arXiv preprint arXiv:2306.07479*, 2023.

Kirthevasan Kandasamy, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Vcg mechanism design with unknown agent values under stochastic bandit feedback. *Journal of Machine Learning Research*, 24(53):1–45, 2023.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Yang Liu and Chien-Ju Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Hamid Nazerzadeh, Renato Paes Leme, Afshin Rostamizadeh, and Umar Syed. Where to sell: Simulating auctions from learning algorithms. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 597–598, 2016.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 129–140, 1999.

David C Parkes. Online mechanisms. 2007.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A myersonian approach. *Econometrica*, 82(2):601–653, 2014.

Philip J Reny. On the existence of pure and mixed strategy nash equilibria in discontinuous games. *Econometrica*, 67(5):1029–1056, 1999.

Suho Shin, Seungjoon Lee, and Jungseul Ok. Multi-armed bandit algorithm against strategic replication. In *International Conference on Artificial Intelligence and Statistics*, pages 403–431. PMLR, 2022.

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

Huazheng Wang, Qingyun Wu, and Hongning Wang. Factorization bandits for interactive recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1288–1297, 2021.

Youtube. How to earn money on YouTube, 2023.

Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*, pages 1011–1018, 2010.

Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.

Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.