# Multi-armed Bandit Mechanism With Private Histories

Chang Liu
Alibaba Group
chang.liu181@gmail.com

Qingpeng Cai
Alibaba Group
cqp14@mails.tsinghua.edu.cn

Yukui Zhang
Alibaba Group
yukui.zyk@alibaba-inc.com

## ABSTRACT

We consider a multi-armed bandit setting in e-commence recommendation, each arm is a new item owned by a seller. In each round, the platform selects one of the items to be delivered to an anonymous user. The user may purchase this item or not. The conversion rate of every item is an unknown parameter both for the platform and the seller. However, the sellers may have some delivered histories of the items, i.e, her private information, and these histories could be used by the platform to do better in minimizing the regret. We present a multi-armed bandits with private histories (PH-MAB) model tailored for this problem, in this model each seller called agent reports her history to the platform called designer. In each round, the designer solicits the reports and the record of pulling arms in previous rounds, outputting an arm selection rule and a payment rule for agents. We study mechanisms that incentivize agents to truthfully report histories of success events to an $\epsilon$-greedy MAB algorithm. We also do experiments on `Taobao` data demonstrating the significant improvements on minimizing regret and increasing revenue of the platform over $\epsilon$-greedy algorithm.

## Keywords

Multi-armed bandit, Mechanism design, Truthful mechanism, Private history, $\epsilon$-greedy

## 1. INTRODUCTION

The fundamental challenge in bandit problem is the elaborate balance between exploration and exploitation. To minimize the regret in a long period, an algorithm has to explore by actually choosing seemingly suboptimal arms so as to gather more information about them. The explorations obviously have higher short-term regrets. In new items recommendation, the lifecycle of these items are remarkably short. We try to gather information as plenty as possible in an exploration process and expect we can get rewards in the following exploitation, but the gains are tiny and some newer items come in and next exploration should be start. We must increase the intensity of exploration so as to gather information quickly, but this will draw more regrets.

**Appears in:** *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*


### 1.1 Our approach

To handle the problem of lack of information for shortly-lived items, we present a novel solution, which not only gathers information from the exploration in the early rounds, but also makes use of history information preserved by agents. Fortunately, it is possible under some circumstances. For example, the sellers may exhibit their items via other channels outside the platform. Also the items may be delivered in a vertical field before published in a complex area.

To involve these information and use them to boost the exploration process will extremely reduce the regret in the earlier rounds as well as the total regret. We should incentivize sellers to report their information truthfully. In this paper we present a multi-armed bandits with private histories (PH-MAB) model that combines mechanism design and multi-arm bandit algorithm. In this model each seller called agent preserves a private history, and reports their histories to the designer. In each round the mechanism solicits the reports and the record of pulling arms in previous rounds, outputting a randomized arm selection rule. Correspondingly, the platform gets a reward from the selected arm and pays the agents. By the well-known revelation principle[15], it is without loss of generality to consider only truthful mechanisms. This model can be viewed as a variant of resource allocation setting[19], and the arm selection rule can be considered as an allocation rule. However, it is different from the traditional model in mainly two aspects. Firstly the utility of an agent is the expected payment she receives conditioned on her private history, which not equals the expected value minus the payment to the designer. Secondly the payment of each round in our setting is not only decided by the reports, but also by the random rewards that agents and planner will observe.

Then we define the consistency, a mechanism is consistent with a MAB algorithm means that the selection rule of arms in the mechanism is the same as the algorithm. We focus on designing truthful mechanisms that are consistent with $\epsilon$-greedy algorithm maximizing the revenue. The revenue equals the sum of rewards minus the sum of payments.

### 1.2 Our contributions

We firstly give two characterizations of truthful mechanisms. With these intuitions, we present a class of truthful-in-expectation single round mechanisms that are consistent with $\epsilon$-greedy algorithm and a class of truthful-in-expectation multi-round mechanisms that are consistent with $\epsilon$-greedy algorithm when the sizes of private history are the same. We simulate the multi-round mechanism in which sellers telling

the truth on real data provided by Taobao, and the result shows that there are significant improvements on the regret minimization and revenue maximization of the platform compared with $\epsilon$-greedy algorithm. Since the sellers may be irrational, to show the robustness of our mechanisms, we implement the mechanism and test on real sellers, finding that there are only 5 precent sellers whose reports deviate far from the truth. According to this, we also simulate the multi-round mechanism in which there are 5 percent sellers misreporting, it turns out that the revenue in this case is slightly less than the revenue with all sellers reporting truthfully, but still outperforms $\epsilon$-greedy algorithm.

# 2. RELATED WORKS

The multi-armed bandit(MAB) setting is one of the most well-known instances of sequential decision-making problems in uncertain environments. It was first proposed by Robbins[18] as a kind of reinforcement learning. The Gittins index theorem[7] gives an optimal policy for maximizing the reward. In the past decade, MAB has widely been applied in many fields such as search engine[17], online advertising[5] and recommender system[12]. It was demonstrated to be effective in solving item cold-start problem in web-based scenarios.

Recently, a rich literature proposed for the bandit problem, some popular ones include upper confidence bound (UCB)[11], which has good theoretical guarantees, and Thompson sampling[4], which performs reliably in quickly changed scenarios. One of the simplest and most straightforward MAB algorithms is $\epsilon$-greedy. It has a very important theoretical significance in the research of MAB. In each round $t$, it first estimates the expected reward of each arm. Then, with probability $1-\epsilon$, it chooses the greedy arm (i.e., the arm with highest expected reward); with probability $\epsilon$, it chooses a random arm. In the limit, each arm will be tried infinitely often, and so the expected reward converges to the true value with probability 1. Some improvements of $\epsilon$-greedy work on make $\epsilon$ shrank gradually, that will make more exploitation and less exploration with time going forward. Auer et al. demonstrate that $O(logn)$ regret can be achieved uniformly over time[1].

The area of mechanism design, which has its origins in the works of Maskin [14] and Myerson [16] is preventive. In this field, the rules of a system are designed in a way that selfish agents will behave as expected.

Several researchers have combined mechanism and MAB in their studies. Adish and Andreas [20] studied on the labour market problem, considering the employer and workers as designer and agents respectively. They proposed a posted-price mechanism, according to which the designer provides adaptive salaries in order to gain the maximized utility. The major point of their study is that, although the true costs of the workers are desirable, the mechanism doesn't explicitly solicit them from the workers. As a result, though remitted by the mechanism design, the truthfulness still affects the ultimate performance. Chien-Ju Ho et al. [9] also focused on the labour market scenario, and treated it as a MAB problem. They considered the contracts between workers and employer as multiple arms, and refined the contract space iteratively. Frazier et al. [6] designed a mechanism, by which the designer offers monetary incentives to encourage the agents to explore more information instead of just to exploit. Specially, they took the time-discounted effect of money into consideration. Li Han et al. [8] also studied the incentive mechanism, while taking into account the heterogeneity of distinguished agents, i.e., the same amount of money is of distinct utilities for different agents. Further discussions on the mechanism design based on the MAB formulation are given in [10, 13, 2, 3].

# 3. MODEL

In this section, we define the multi-armed bandit with private histories model formally. Before it, we revisit the MAB setting at first.

## 3.1 Setting of multi-armed bandit

Let there be $K$ arms, $[K] = \{1, 2, ..., K\}$. The reward of arm $k$, $r_k$ is drawn from a probability distribution, with density function , $f(x, \theta_k)$, where $\theta_k$ is a vector of deterministic unknown parameters. The prior distributions of all parameters are i.i.d, and denoted by $D$. In this paper we assume that $r_k$ is drawn from a Bernoulli distribution $B(1, \theta_k)$, and the prior distribution $D$ is a Beta distribution, $Beta(1, 1)$. At each round $t = 1, 2, ..., T$, the designer selects an arm $a_t \in [K]$ to pull and then receives the reward $r_t(a_t)$. After all $T$ rounds, the total reward is, $Rew(T) = \sum_{t=1}^{T} r_t(a_t)$. Without loss of generality, we assume that $\mu_{(1)}$ is the maximum expected reward among all arms. We can also use regret to measure the difference between an ideal best algorithm and the actual one, i.e, $Reg(T) = T\mu_{(1)} - Rew(T)$. The objective of this problem is minimizing the regret.

## 3.2 PH-MAB Model

In this section, we define a multi-armed bandits with private histories (PH-MAB) model. In the model, there are two stages. In the first stage, every agent collects his own pulling history as private information, we call these Private History(PH). At the end of first stage, every agent is demanded to report their PH. In the second stage, the platform operate the multi-armed machine and finally allocate the utility. In this stage, the pulling results can be observed by all the agents and the platform, so as we call these results Common History(CH). More specifically, there are $K$ agents, the PH is the rewards of each pulling round in first stage. Let $H_k = \{r_{k,t}\}$ denote the history of agent $k$($r_{k,t}$ means the reward of arm $k$ of $t-$th pulling in the first stage), $H_k^{'}$ denote the reported history of agent $k$ and $H_{-k}^{'}$ denote the reported histories of agents except from agent $k$. Also, $H = (H_1, ..., H_K)$ and $H^{'} = (H_1^{'}, ..., H_K^{'})$.

DEFINITION 1. *A multi-armed bandits with private histories (PH-MAB) mechanism $f$ is a process of $T$ rounds:*

- *Before round 1, each agent k reports their private pulling history of his own arm to the designer, i.e, $H_k^{'}$.*
- *At each round $t(1 \le t \le T)$, the mechanism $f$ solicits the reports $H^{'}$, and the records of the pulling arm and the rewards in previous rounds, denoted by $R_{t-1}$, outputs a randomized rule selecting the arm with probability distribution $f(t, H^{'}, R_{t-1})$.*

  *After receiving the reward of the arm $r_t(a_t)$, each agent $k$ receives the payment from the mechanism*

  $p(k, t, H^{'}, R_{t-1}, r_t(a_t))$.

The utility of each agent $k$ with a PH-MAB mechanism $f$ and a report $H^{'}$ is the expectation of the sum of payments received from the mechanism, conditioned on the history of agent $k$,

$$u_k(f, H_k, H^{'}) = \sum_{t=1}^{T} E[p(k, t, H^{'}, R_{t-1}, r_t(a_t))|H_k]. \quad (1)$$

The definition is conditioned on $H_k$, rather than $\theta_k$. It is because that $\theta_k$ is an unknown parameter for the agent, $H_k$ is her only observation.

There are several desirdata that a PH-MAB mechanism must satisfy.

DEFINITION 2. *IR: A mechanism $f$ is individual rational if for any agent $k$, the utility of agent $k$ is nonnegative with truthfully reporting his history and other agents reporting arbitrarily, i.e, $\forall k \forall H_k \forall H^{'}_{-k}, u_k(f, H_k, (H_k, H^{'}_{-k})) \geq 0$.*

DEFINITION 3. *IC: A mechanism is truthful-in-expectation if for any agent $k$, with any true history $H_k$, and other agents reporting $\{H'_{-k}\}$ arbitrarily, any misreporting $H'_k$ will lead to no more expected utility, i.e,*

$$\forall k \forall H_k \forall H^{'}_k \forall H^{'}_{-k}, u_k(f, H_k, (H_k, H^{'}_{-k})) \geq u_k(f, H_k, (H^{'}_k, H^{'}_{-k})). \quad (2)$$

The reason why agents report their history is that we can use the truthful prior history to do better in minimizing the regret combined with the existing MAB algorithms in the literature, such as $\epsilon-$ greedy, Thompson sampling and UCB. Now we define the consistency.

DEFINITION 4. *CONSISTENCY: A PH-MAB mechanism $f$ is consistent with a MAB algorithm $g$ if for any reported histories of all agents and any pulling histories of previous rounds, $f$ and $g$ use the same randomized rule selecting the arm, i.e, $\forall t \forall H^{'} \forall R_{t-1}, f(t, H^{'}, R_{t-1}) = g(t, H^{'}, R_{t-1})$.*

Our objective is to design truthful, individual rational mechanism which is consistent with $\epsilon-$ greedy algorithm maximizing the revenue of the designer. Here we define the revenue.

DEFINITION 5. *The revenue of a PH-MAB mechanism $f$, given the truthfully reports $H$, is $Rev(f, H) = \sum_{t=1}^{T} E[r_t|f, H] - \sum_{k=1}^{K} u_k(f, H)$.*

The revenue of the designer equals the summation of the expectation of rewards of $T$ rounds minus the sum of the payments to agents conditioned on the reports of history. Also, we define the regret.

DEFINITION 6. *Given a truthfully report $H$, the regret of a PH-MAB mechanism $f$ is $Reg(f, H) = T\mu_{(1)} - \sum_{t=1}^{T} E[r_t|f, H]$.*

# 4. OUR RESULT

## 4.1 Intuitions and characterizations

Before presenting our results, we first show characterizations of truthful mechanisms.

Given a history of agent $i$, $H_i$, the posterior distribution of $\theta_i$ is $p(\theta_i|H_i)$. Then we can calculate the expected reward for arm $i$ conditioned on the history $H_i$, denoted by $R(H_i)$,

$$R(H_i) = E[r_i|H_i] = \int E[r_i|\theta_i] p(\theta_i|H_i) \mathrm{d}\theta_i \quad (3)$$

In particular, let $M(H_i)$ be the number of 1 in the set $H_i$, and $S(H_i)$ be the size of the set $H_i$. By the property of Beta distributions, $R(H_i) = \frac{M(H_i)+1}{S(H_i)+2}$.

Similarly, Let $R(H_{-i})$ denote the maximum of the expected reward of other arms conditioned on their own history except $i$, i.e, $R(H_{-i}) = Max_{j \neq i} R(H_j)$. By the analysis of the truthfulness, we get the following theorems.

THEOREM 1. *For any single round truthful mechanism of which the payment function is affine with the reward , the utility of agent $k$ telling the truth is determined by the mechanism, the reports of other agents and the expected reward for arm $k$ conditioned on the history $H_k$, i.e,*

$$\forall k \forall f \forall H^{'}_{-k} \forall H^1_k \forall H^2_k (R(H^1_k) = R(H^2_k)),$$
$$u_k(f, H^1_k, (H^1_k, H^{'}_{-k})) = u_k(f, H^2_k, (H^2_k, H^{'}_{-k})).$$

PROOF. For any agent $k$, and any two pulling history $H^1_k, H^2_k$ with the same conditional expected reward, by the definition of the utility, we get that

$$u_k(f, H^1_k, (H^1_k, H^{'}_{-k})) = E[p(k, 1, (H^1_k, H^{'}_{-k}), R_0, r_t(a_t))|H^1_k]. \quad (4)$$

$$u_k(f, H^2_k, (H^2_k, H^{'}_{-k})) = E[p(k, 1, (H^2_k, H^{'}_{-k}), R_0, r_t(a_t))|H^2_k]. \quad (5)$$

Then we consider agent $k$ with real type $H^1_k$ misreporting his history as $H^2_k$, by the truthfulness,

$$u_k(f, H^1_k, (H^1_k, H^{'}_{-k})) \geq u_k(f, H^1_k, (H^2_k, H^{'}_{-k})). \quad (6)$$

and

$$u_k(f, H^1_k, (H^2_k, H^{'}_{-k})) = E[p(k, 1, (H^2_k, H^{'}_{-k}), R_0, r_t(a_t))|H^1_k]. \quad (7)$$

Comparing equations (5) with (7), the reports are the same, thus the arm selected by the mechanism is the same in two cases, the expected rewards are the same, and the payment function $p$ is affine with the reward $r_t(a_t)$, then

$$u_k(f, H^1_k, (H^2_k, H^{'}_{-k})) = u_k(f, H^2_k, (H^2_k, H^{'}_{-k})). \quad (8)$$

Combining (6) with (8),

$$u_k(f, H^1_k, (H^1_k, H^{'}_{-k})) \geq u_k(f, H^2_k, (H^2_k, H^{'}_{-k})). \quad (9)$$

By the same approach, we get

$$u_k(f, H^2_k, (H^2_k, H^{'}_{-k})) \geq u_k(f, H^1_k, (H^1_k, H^{'}_{-k})). \quad (10)$$

Combining (9) with (10), the theorem holds. □

The above theorem shows that when we design an affine payment function for each agent $k$, we need to only consider $R(H_k)$, regardless other characterizations of the history $H_k$. The affinity is a common character for payment function, for example, when the reward of every arm is binary, the payment function is definitely affine with the reward.

For single round truthful mechanism consistent with $\epsilon$-greedy($\epsilon = 0$) mechanism, we get the following theorem.

THEOREM 2. *For any single round truthful mechanism, the utility of any agent whose conditional expected reward is not highest among the conditional expected rewards of reports is independent of his own conditional expected reward, i.e,*

$$\forall k \forall H^1_k \forall H^2_k \forall H^{'}_{-k} (R(H^1_k), R(H^2_k) \leq R(H^{'}_{-k}))$$
$$u_k(f, H^1_k, (H^1_k, H^{'}_{-k})) = u_k(f, H^2_k, (H^2_k, H^{'}_{-k})).$$

PROOF. *For any agent $k$ whose conditional expected reward is not highest among the conditional expected rewards of reports and any truthful mechanism $f$, let $i = argmax_{j\neq k}R(H_j)$. By the definition of the utility,*

$$u_k(f, H_k^1, (H_k^1, H_{-k}^{'})) = E[p(k, 1, (H_k^1, H_{-k}^{'}), f, R_0, r_1(i))|H_k^1]. \tag{11}$$

$$u_k(f, H_k^2, (H_k^2, H_{-k}^{'})) = E[p(k, 1, (H_k^2, H_{-k}^{'}), f, R_0, r_1(i))|H_k^2]. \tag{12}$$

*By the truthfulness,*

$$u_k(f, H_k^1, (H_k^1, H_{-k}^{'})) \geq u_k(f, H_k^1, (H_k^2, H_{-k}^{'})). \tag{13}$$

*and*

$$u_k(f, H_k^1, (H_k^2, H_{-k}^{'})) = E[p(k, 1, (H_k^2, H_{-k}^{'}), f, R_0, r_1(i))|H_k^1]. \tag{14}$$

*Because of the independence of $r_1(i)$ and $H_k^1$, $H_k^2$,*

$$\begin{aligned} &E[p(k, 1, (H_k^2, H_{-k}^{'}), f, R_0, r_1(i))|H_k^2] \\ &= E[p(k, 1, (H_k^2, H_{-k}^{'}), f, R_0, r_1(i))|H_k^1]. \end{aligned} \tag{15}$$

*Combing all above formulas,*

$$u_k(f, H_k^1, (H_k^1, H_{-k}^{'})) \geq u_k(f, H_k^2, (H_k^2, H_{-k}^{'})). \tag{16}$$

*By the same way, we can get*

$$u_k(f, H_k^2, (H_k^2, H_{-k}^{'})) \geq u_k(f, H_k^1, (H_k^1, H_{-k}^{'})). \tag{17}$$

*Combing (16) with (17), the theorem holds.* □

With the guidance of these two theorems, we have the following intuition in designing mechanism consistent with $\epsilon$-greedy algorithm. There are two cases, firstly, the agent's arm is selected in this round, secondly, it is not selected. In the first case, the payment should consist of two parts: (1)the reward $r_t$ (2)The increments of the social benefit increased over other arms $r_t - R(H_{-i})$. These 2 parts should be assigned proper weights to form the payments to agents. In the second case, the arm of the agent is not selected, so we pay her a proportion from the expected reward from the best arm which is independent of her own history.

## 4.2 A class of single round mechanism

In the section, we present two classes of single round mechanisms, one is consistent with $\epsilon$-greedy ($\epsilon = 0$) MAB mechanism, the other is consistent with $\epsilon$-greedy ($\epsilon > 0$) MAB mechanism, and prove that they are truthful.

MECHANISM 1. *Arm selection: Let $i = argmax_k(R(H_k^{'}))$, the mechanism selects arm $i$. Payments:*

$$p(k, 1, H^{'}, f, R_0) = \begin{cases} \lambda_1 r_1(i) - \lambda_2 R(H_{-i}^{'}) & k = i \\ \lambda_3 R(H_i^{'}) & k \neq i \end{cases} \tag{18}$$

*$\lambda_1, \lambda_2, \lambda_3$ are nonnegative parameters, $\lambda_1 = \lambda_2 + \lambda_3$, and $R_0$ is an empty set.*

Obviously Mechanism 1 is consistent with $\epsilon$-greedy ($\epsilon = 0$) MAB algorithm. By the definition, $\lambda_1 \geq \lambda_2$, which ensures that Mechanism 1 is IR. If we choose $\lambda_3 = 0$, and $\lambda_1 = \lambda_2 = 1$, the expected utility of each agent is the same as it in second price auction in which each agent $k$ has type $R(H_k)$. We can prove that this class of mechanisms are all truthful.

THEOREM 3. *Mechanism 1 is truthful.*

PROOF. The selection of the arm and the payments are determined by the largest value and the second largest value in $R(H^{'}) = (R(H_1^{'}), ..., R(H_K^{'}))$. We consider two major cases in which the selection of the arm and the payments will change with the misreport.

**Case 1:** Misreporting an inferior arm to the best : $R(H_k') > R(H_{-k}') > R(H_k)$ Let $j = argmax_{i,i\neq k}R(H_i')$. Thus $R(H_{-k}') = R(H_j')$. Given the report $(H_k, H_{-k}^{'})$ the mechanism selects arm $j$, the utility of agent $k$ is $u_k(f, H_k, (H_k, H_{-k}^{'})) = \lambda_3 R(H_j')$.

And given the report $(H_k^{'}, H_{-k}^{'})$ the mechanism selects arm $k$, the utility of agent $k$ is

$$\begin{aligned} &u_k(f, H_k, (H_k^{'}, H_{-k}^{'})) = E[\lambda_1 r_1(k)|H_k] - \lambda_2 R(H_{-k}') \\ &= \lambda_1 R(H_k) - \lambda_2 R(H_j'). \end{aligned} \tag{19}$$

Then

$$\begin{aligned} &u_k(f, H_k, (H_k^{'}, H_{-k}^{'})) - u_k(f, H_k, (H_k, H_{-k}^{'})) \\ &= \lambda_1(R(H_k) - R(H_j')) \leq 0. \end{aligned} \tag{20}$$

**Case 2:** Misreporting the best arm to be an inferior: $R(H_k) > R(H_{-k}') > R(H_k')$ Let $j = argmax_{i,i\neq k}R(H_i')$. Given the report $(H_k, H_{-k}^{'})$ the mechanism selects arm $k$, the utility of agent $k$ is

$$u_k(f, H_k, (H_k, H_{-k}^{'})) = E[\lambda_1 r_1(k)|H_k] - \lambda_2 R(H_{-k}') \tag{21}$$

$$= \lambda_1 R(H_k) - \lambda_2 R(H_j'). \tag{22}$$

And given the report $(H_k^{'}, H_{-k}^{'})$ the mechanism selects arm $j$, the utility of agent $k$ is $u_k(f, H_k, (H_k^{'}, H_{-k}^{'})) = \lambda_3 R(H_j')$. Then

$$\begin{aligned} &u_k(f, H_k, (H_k^{'}, H_{-k}^{'})) - u_k(f, H_k, (H_k, H_{-k}^{'})) \\ &= \lambda_1(R(H_j') - R(H_k)) \leq 0. \end{aligned} \tag{23}$$

In all, telling the truth is a dominant strategy for each agent. □

We use the same intuition to design a mechanism consistent with $\epsilon$-greedy algorithm.

MECHANISM 2. *We first generate a random variable $I$ drawn from a Bernoulli Distribution $B(1, \epsilon)$. Arm selection:*

$$a_t = \begin{cases} uniformly - selection & I = 1 \\ argmax_k(R(H_k^{'})) & I = 0 \end{cases} \tag{24}$$

*Payments: Let $i = argmax_k(R(H_k^{'}))$.*

$$p(k, 1, H^{'}, f, R_0) = \begin{cases} \lambda_4 r_1 & I = 1 \\ \lambda_1 r_1(i) - \lambda_2 R(H_{-i}^{'}) & k = i \& I = 0 \\ \lambda_3 R(H_i^{'}) & k \neq i \& I = 0 \end{cases} \tag{25}$$

*$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are nonnegative parameters, $\lambda_1 = \lambda_2 + \lambda_3$.*

It is easy to see Mechanism 2 is consistent with $\epsilon$-greedy MAB algorithm. If $I$ equals to 1, the reports of agents will not affect the payments of all agents, else the payments is the same as the payments of Mechanism 1. Thus we get the following theorem and we omit the proof because it is the same as the proof of Theorem 3.

THEOREM 4. *Mechanism 2 is truthful.*

## 4.3 Multi-round mechanism

In this section we present a multi-round mechanism using the idea of Mechanism 2. Set the record of arm $k$ before round $t$ as $R_{k,t} = \{r_{t'} | a_{t'} = k, t' < t\}$ and the total histories before round $t$ as $H^{all}_{k,t} = \{H^{'}_{k}, R_{k,t}\}$.

MECHANISM 3. *We first generate a random variable $I$ drawn from a Bernoulli Distribution $B(1,\epsilon)$. Arm selection:*

$$a_t = \begin{cases} randomly - selection & I = 1 \\ argmax_k(R(H^{all}_{k,t})) & I = 0 \end{cases} \tag{26}$$

*Payments: Let $i = argmax_k(R(H^{all}_{k,t}))$,*

$$p(k,t,H^{'},f,R_{t-1}) = \begin{cases} \lambda_4 r_t & I = 1 \\ \lambda_1 r_t - \lambda_2 R(H^{all}_{-i,t}) & k = i \& I = 0 \\ \lambda_3 R(H^{all}_{i,t}) & k \neq i \& I = 0 \end{cases} \tag{27}$$

*$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are nonnegative parameters, $\lambda_1 = \lambda_2 + \lambda_3$.*

We combine the reports of agents $H^{'}_{k}$(Private History) and the record $R_{k,t}$(Common History) as the whole history used for arm selection and payment calculation. We now prove that Mechanism 3 is truthful when the size of each seller's private history is the same and such is the size of reported history.

THEOREM 5. *Mechanism 3 is truthful when the size of each seller's private history and the size of reported history of each seller are the same.*

PROOF. Without loss of generality, we consider the case that $\epsilon = 0$ and $T = 2$. For any agent $k$, we prove that the dominant strategy for this agent is to tell the truth with any reported history of other agents. We prove it in four cases.
**Case 1:** the conditional expected reward of agent $k$ is highest and he misreports a history with higher expected reward, i.e, $R(H^{'}_{k}) > R(H_k) > R(H^{'}_{-k})$. Given other agents reporting $H^{'}_{-k}$, the utility of agent $k$ telling the truth is

$$\begin{aligned} u_k(f, H_k, (H_k, H^{'}_{-k})) &= \lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k}) \\ &+ p(r_1 = 1|H_k)(\lambda_1 R(H_k \cup 1) - \lambda_2 R(H^{'}_{-k})) \\ &+ p(r_1 = 0|H_k) \\ &\begin{cases} \lambda_1 R(H_k \cup 0) - \lambda_2 R(H^{'}_{-k}) & R(H_k \cup 0) > R(H^{'}_{-k}) \\ \lambda_3 R(H^{'}_{-k}) & R(H_k \cup 0) < R(H^{'}_{-k}) \end{cases} \end{aligned} \tag{28}$$

Then we can get that

$$u_k(f, H_k, (H_k, H^{'}_{-k})) \geq 2(\lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k})). \tag{29}$$

As seller $k$ misreports a history with higher expected reward, Mechanism 3 will still select arm $k$ in round 1. In round 2 if $r_1 = 1$ Mechanism 3 selects arm $k$ because $R(H^{'}_{k} \cup 1) > R(H_k \cup 1)$, else if $r_1 = 0$ and $R(H_k \cup 0) > R(H^{'}_{-k})$ Mechanism 3 still selects arm $k$ because $R(H^{'}_{k} \cup 0) > R(H_k \cup 0)$, else Mechanism 3 selects arm $k$ and in the case that agent tells the truth Mechanism 3 selects arm $-k$. Thus misreporting a history with higher expected reward may decrease the seller's utility because $\lambda_3 R(H^{'}_{-k}) \geq \lambda_1 R(H_k \cup 0) - \lambda_2 R(H^{'}_{-k})$.
**Case 2:** the conditional expected reward of agent $k$ is highest and he misreports a history with lower expected reward, i.e, $R(H_k) > R(H^{'}_{-k}) > R(H^{'}_{k})$. Note that it is easy to see that the utility of seller $k$ decreases in the case that $R(H_k) > R(H^{'}_{k}) > R(H^{'}_{-k})$. Then the utility of seller $k$ telling $H^{'}_{k}$ is

$$\begin{aligned} u_k(f, H_k, (H^{'}_{k}, H^{'}_{-k})) &= \lambda_3 R(H^{'}_{-k}) + p(r_1 = 1)\lambda_3 R(H^{'}_{-k} \cup 1) \\ &+ p(r_1 = 0) \begin{cases} \lambda_3 R(H^{'}_{-k} \cup 0) & R(H^{'}_{-k} \cup 0) > R(H^{'}_{k}) \\ \lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k} \cup 0) & R(H^{'}_{-k} \cup 0) < R(H^{'}_{k}) \end{cases} \end{aligned} \tag{30}$$

Thus

$$\begin{aligned} &u_k(f, H_k, (H^{'}_{k}, H^{'}_{-k})) \leq \\ &max(\lambda_3 R(H^{'}_{-k} \cup 1), \lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k} \cup 0)) + \lambda_3 R(H^{'}_{-k}). \end{aligned} \tag{31}$$

Let $S(H_k) = S(H^{'}_{-k}) = a$, then

$$2(\lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k})) = \frac{2(\lambda_1(M(H_k)+1) - \lambda_2(M(H^{'}_{-k})+1))}{a+2}. \tag{32}$$

$$\lambda_3 R(H^{'}_{-k} \cup 1) + \lambda_3 R(H^{'}_{-k}) = \lambda_3(\frac{M(H^{'}_{-k})+2}{a+3} + \frac{M(H^{'}_{-k})+1}{a+2}). \tag{33}$$

$$\begin{aligned} &\lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k} \cup 0) + \lambda_3 R(H^{'}_{-k}) \\ &= \lambda_1 \frac{M(H_k)+1}{a+2} - \lambda_2 \frac{M(H^{'}_{-k})+1}{a+3} + \lambda_3 \frac{M(H_{-k})+1}{a+2}. \end{aligned} \tag{34}$$

Combining with formulas (29)-(34), we can prove that

$$u_k(f, H_k, (H^{'}_{k}, H^{'}_{-k})) \leq u_k(f, H_k, (H_k, H^{'}_{-k})).$$

**Case 3:** the conditional expected reward of agent $k$ is not highest and he misreports a history with lower expected reward, i.e, $R(H^{'}_{-k}) > R(H_k) > R(H^{'}_{k})$. the utility of seller $k$ telling the truth is

$$\begin{aligned} u_k(f, H_k, (H_k, H^{'}_{-k})) &= \lambda_3 R(H^{'}_{-k}) + p(r_1 = 1)\lambda_3 R(H^{'}_{-k} \cup 1) \\ &+ p(r_1 = 0) \begin{cases} \lambda_3 R(H^{'}_{-k} \cup 0) & R(H^{'}_{-k} \cup 0) > R(H_k) \\ \lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k} \cup 0) & R(H^{'}_{-k} \cup 0) < R(H_k) \end{cases} \end{aligned} \tag{35}$$

It is easy to see misreporting a history with lower expected reward will decrease seller's utility using the analysis similar to it in case 1.
**Case 4:** the conditional expected reward of agent $k$ is not highest and he misreports a history with highest expected reward, i.e, $R(H^{'}_{k}) > R(H^{'}_{-k}) > R(H_k)$. It is easy to see that misreporting a history with higher expected reward but not highest will decrease the seller's utility. From formula (35),

$$u_k(f, H_k, (H_k, H^{'}_{-k})) \geq \lambda_3 R(H^{'}_{-k}) + \lambda_3 R(H^{'}_{-k} \cup 0). \tag{36}$$

And we can get that

$$\begin{aligned} &u_k(f, H_k, (H^{'}_{k}, H^{'}_{-k})) \leq \\ &\lambda_1 R(H_k) - \lambda_2 R(H^{'}_{-k}) + \lambda_1 R(H_k \cup 1) - \lambda_2 R(H^{'}_{-k}). \end{aligned} \tag{37}$$

Combining with (36) and (37), we can prove that

$$u_k(f, H_k, (H^{'}_{k}, H^{'}_{-k})) \leq u_k(f, H_k, (H_k, H^{'}_{-k})) \tag{38}$$

□

# 5. EXPERIMENTS

Taobao App is one of the most popular online shopping sites around the world. There are hundreds of thousands new items submitted into the platform by different sellers everyday. There is a scenario named *Daily New Goods* in Taobao App's homepage to exhibit and sale these items. In this scenario, we have a limited opportunity to show these items everyday, and what we concern is the total transactions. So we try to find the best or similar items with the highest conversion rate, and to give them more opportunity to be exhibited. The platform generally uses a MAB algorithm on different human groups to deliver different items. In this section, we simplify the problem as a pure MAB algorithm on single human group. We regard the seller of every item is an agent, also we suppose that every seller has only one item. We use the data in this scenario to implement the following experiments.

## 5.1 An example of Mechanism 3

Before showing the experiment results, we first display an example to explain how Mechanism 3 works. In this example, we assume there are 3 agents, and their reporting private histories are showing in the first row of Table.1. The designer run a 10 rounds $\epsilon - greedy(\epsilon = 0.1)$ mechanism. The reward and utility of every agent is also in the table. Notice that, in every round except the seventh one, the arm with the maximum expected value is selected to pull(The tie-breaker sequences is 1,2,3). While in the seventh round, the 0.1 incidence happened and the arm is randomly selected. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.008, \lambda_3 = 0.002, \lambda_4 = 0$ in this experiment. To summarize the experiment results, Agent-1 get 4 times to be selected and 1 positive result from that, Agent-2 get 5 times to be selected and 4 positive results from that, while Agent-3 get 1 positive result from $\epsilon$ incidence. The total utilities of the three agents are $-3.23 \times 10^{-3}$, $22.17 \times 10^{-3}$ and $12.28 \times 10^{-3}$ respectively. Agent-1 get a negative utility mainly caused of the inconsistence between PH $\{1, 1, 0, 1, 1, 1\}$ and CH $\{0, 1, 0, 0\}$. The underlying reason of this result may be the untruthfulness of Agent-1 or just a bad luck of Agent-1 in the first 4 rounds.

## 5.2 The performance of Mechanism 3

To estimate the performance, we focus on two objectives. The first one is regret, an index of the loss of social welfare including platform and all the sellers. The other is the revenue of platform. We assume the probability of an item being conversed is $\mu_k \in (0, 1)$, which is deterministic, but unknown. Hence the reward of selecting this arm is under a Bernoulli distribution $r_k \sim B(1, \mu_k)$. We select 10 popular items from the *Daily New Goods* scenario. Their frequency of exposures in one day are over 10 thousands. Because the parameters of items are unknown, we use the posteriori estimation $\frac{\sharp transaction}{\sharp exposure}$ to be $\mu_k$. We assume the first 50 exposures are private information of the sellers. In our implementation, there are four mechanisms to be compared:

1. $\epsilon$-greedy with $\epsilon = 0.02$ (Ignore the private information, the agents(sellers) are not involved in the mechanism, do exploration from the virgin paper)
2. $\epsilon$-greedy with $\epsilon = 0.01$
3. Mechanism 3($\epsilon = 0.02$)
4. Mechanism 3($\epsilon = 0.01$)

**Table 1: An example of Mechanism 3**

| PH | {1,1,0,1,1,1} | {1,1,1,0,1,0} | {0,1,0,1,0,0} |
|---|---|---|---|
| Agent | Agent-1 | Agent-2 | Agent-3 |
| Round 1 | | | |
| EV | 0.75 | 0.625 | 0.375 |
| AS | Yes(0) | No | No |
| Ut | $0\lambda_1$ $-0.625\lambda_2$ | $0.75\lambda_3$ | $0.75\lambda_3$ |
| Round 2 | | | |
| EV | 0.667 | 0.625 | 0.375 |
| AS | Yes(1) | No | No |
| Ut | $1\lambda_1$ $-0.625\lambda_2$ | $0.667\lambda_3$ | $0.667\lambda_3$ |
| Round 3 | | | |
| EV | 0.7 | 0.625 | 0.375 |
| AS | Yes(0) | No | No |
| Ut | $0\lambda_1$ $-0.625\lambda_2$ | $0.7\lambda_3$ | $0.7\lambda_3$ |
| Round 4 | | | |
| EV | 0.636 | 0.625 | 0.375 |
| AS | Yes(0) | No | No |
| Ut | $0\lambda_1$ $-0.625\lambda_2$ | $0.636\lambda_3$ | $0.636\lambda_3$ |
| Round 5 | | | |
| EV | 0.583 | 0.625 | 0.375 |
| AS | No | Yes(1) | No |
| Ut | $0.625\lambda_3$ | $1\lambda_1$ $-0.583\lambda_2$ | $0.625\lambda_3$ |
| Round 6 | | | |
| EV | 0.583 | 0.667 | 0.375 |
| AS | No | Yes(1) | No |
| Ut | $0.667\lambda_3$ | $1\lambda_1$ $-0.583\lambda_2$ | $0.667\lambda_3$ |
| Round 7 | | | |
| EV | 0.583 | 0.7 | 0.375 |
| AS | rand-No | rand-No | rand-Yes(1) |
| Ut | $1\lambda_4$ | $1\lambda_4$ | $1\lambda_4$ |
| Round 8 | | | |
| EV | 0.583 | 0.7 | 0.444 |
| AS | No | Yes(1) | No |
| Ut | $0.7\lambda_3$ | $1\lambda_1$ $-0.583\lambda_2$ | $0.7\lambda_3$ |
| Round 9 | | | |
| EV | 0.583 | 0.727 | 0.444 |
| AS | No | Yes(0) | No |
| Ut | $0.727\lambda_3$ | $0\lambda_1$ $-0.583\lambda_2$ | $0.727\lambda_3$ |
| Round 10 | | | |
| EV | 0.583 | 0.667 | 0.444 |
| AS | No | Yes(1) | No |
| Ut | $0.667\lambda_3$ | $1\lambda_1$ $-0.583\lambda_2$ | $0.667\lambda_3$ |
| CH | {0,1,0,0} | {1,1,1,0,1} | {1} |
| Total Ut | $-3.23 \times 10^{-3}$ | $22.17 \times 10^{-3}$ | $12.28 \times 10^{-3}$ |
| PR | $6 + 3.23 \times 10^{-3} - 22.17 \times 10^{-3}$ $-12.28 \times 10^{-3} = 5.969$ | | |

**PH:Private Histories;EV:Expect Value of every arm; AS:Arm Selection and $r_t$;Ut:Utility;CH:Common Histories; PR:Platform Revenue**

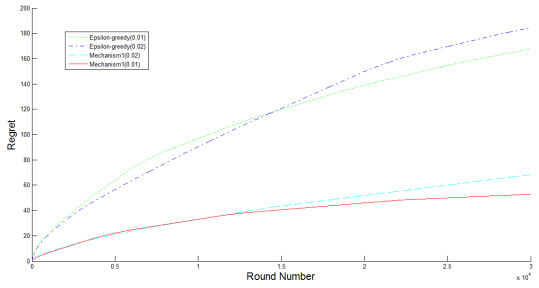


**Figure 1: Comparison of accumulative regret**

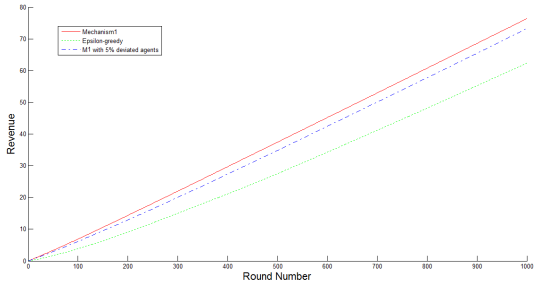


**Figure 2: Comparison of accumulative platform's revenue**

In the *Daily New Goods* scenario, some fixed ratio of immediate reward $r_t$ is directly get by the seller. So $\lambda_1$ has a lower limit greater than zero. We set $\lambda_1 = 0.01$ in this experiment. The other parameters are $\lambda_2 = 0.008, \lambda_3 = 0.002, \lambda_4 = 0$. We plot the accumulative regrets in every round in Fig. 1. We can conclude that the regret of Mechanism 3 is about one third of $\epsilon$-greedy. In the comparison of different $\epsilon$, we can see that larger $\epsilon$ will make the exploration process quicker so as to get less regret in short term, however in the other hand, larger $\epsilon$ lead into more regret in long term. In Mechanism 3, the short term gap between larger and smaller $\epsilon$ is greatly reduced. So we can choose smaller $\epsilon$ to earn the benefit in long term.

Furthermore, we are also interested in the platform's revenue. Because for the platform's perspective, if it can't benefit from the mechanism comparing with only running a pure MAB algorithm, it will have no motivation to run this mechanism. Notice that we test the expected revenus of Mechanism 3, i.e, $E_H Rev(f, H)$. From Fig. 2, we can see that the revenue of platform exceeds $\epsilon$-greedy(in $\epsilon$-greedy, we sum up all the arm results as the platform revenue) significantly. In the end of 1000 rounds, the revenue of Mechanism 3 is 76.32(red line), which is 22.6% greater than that of $\epsilon$-greedy(green line, 62.25).

## 5.3 Will the agents be rational in Mechanism 3?

In the previous subsection, we show that with the truthfully reported information, Mechanism 3 performs significantly better than the traditional MAB algorithm while the platform get considerable revenue. Based on the truthfully reports, the platform can get more information of arms, so as to do high efficiency exploitation and low limited exploration from the beginning. However, the premise of truthfully reporting is the rationality of the agents, it is highly influenced by the information transparency. Actually, the agent may submit truthless information, the reasons include: a)the platform can't explain the detail of the mechanism well, and the sellers has no fully patience to understand them; b) because of the existence of randomness, the agent tend to use domain knowledge to amend the real data, furthermore, randomness can also induce the agents' gambler mentality; c)A strategy against the competitor: even if an agent can't be benefited from a strategy, he may disrupt the market and make the competitor also not benefiting from it. To inspect the unreasonable behavior under Mechanism 3, we did the following psychology experiments.

We implemented Mechanism 3, and tested it on 20 different agents(people) for 30000 rounds. Every agent holds an item with about 500 private histories, these items are picked from the real delivery data in *Daily New Goods* scenario. A brief introduction of Mechanism 3 was shown to the agent in the following text (originally in Chinese), the detail of Mechanism 3 was also shown to all the agent (As expected, most testees didn't pay much attention on this part due to the complicated formulas). To make the experiments brief, we fixed the deliver count and only make the agent to submit conversion count.

*Welcome to this experiment, in this experiment, you are assume to be a seller. You have an item to be sold on our platform, however, your item will be competed with 19 other items. Your income will be consist of three parts. First, once your item is sold, you will get a fixed profit. Second, if the platform don't deliver your item at one time, we will give you a share of the platform revenue. Third, if the delivering results are not significantly exceed the information you submit, we will give you a bonus, otherwise, if the results deflect the information you submit, we will give you some punishments(please note: due to randomness in conversion, we may punish you by fault, please forgive us). You will have 6 chances of simulation before a formal game. You will receive a small gift if you get more than 1000 points revenue in the formal game. Thank you!*

We ran 6 simulation games before the formal game, the testee can learn the mechanism from the simulation games. We summarize (in Table.2) the submission in the simulation games and the formal game in the following way.

- Exact: the submission number is exactly the same as the true number in private history;
- Similar: the submission number is around the private history in a $\pm 15\%$ range;
- Deviated: the submission number is out the range of $\pm 15\%$.

Most of the agents tended to report truthfully at the first. Then they tried different reporting strategies afterward. And finally in the formal game, more than half of the agents were back to report truthfully, and nearly all the agents didn't deviate from the true history to much. On the other hand, we have observed some interesting phenomena, the behaviors of different agents are very different from each other. Some agents chose to report the truth from the beginning to the end, while some others tried their best to

**Table 2: Agents' reports summarization in different stage of game**

| Class | First game | 2nd-6th games | Formal game |
|---|---|---|---|
| `Exact` | 80% | 37% | 55% |
| `Similar` | 0% | 19% | 40% |
| `Deviated` | 20% | 44% | 5% |

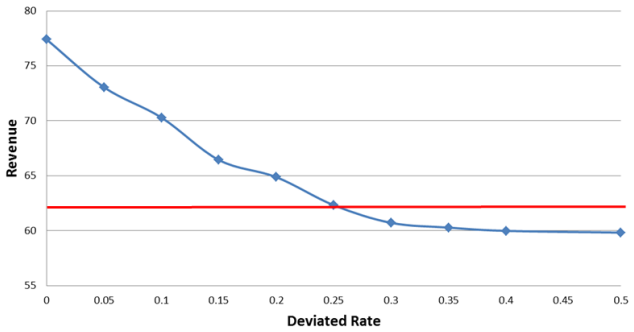


**Figure 3: Platform's revenue in different deviated rate (The horizontal line is the baseline revenue of $\epsilon$-greedy)**

fake their own data, and some individuals' behaviors seem unreasonable.

## 5.4 The robustness of Mechanism 3

We are interested in the robustness of Mechanism 3. In the condition that some irrational sellers exist, does the platform get a similar revenue?

We used the data we get in Sec.5.3 to simulate Mechanism 3 again. From Fig. 2(blue line), we can claim that with 55% `Exact` agents, 40% `Similar` agents, Mechanism 3 get a revenue of 73.18 at the end of 1000 rounds. It works a little weaker than Mechanism 3 with truthfully reports, but outperforms $\epsilon$-greedy in 17.6%.

To further demonstrate the robustness of Mechanism 3, we use different deviated rates to investigate. Deviated rate means the percentage of agents that will report deviated from the truth. We simulated Mechanism 3 with 1000 rounds in different deviated rates from the agents. The revenue of $\epsilon$-greedy is 62.25, and was set as the baseline for Mechanism 3. We can conclude from Fig. 3 that with no more than 25% deviated rate, Mechanism 3 can beat $\epsilon$-greedy with the exceeding revenue.

# 6. FUTURE WORK

As Thompson sampling outperforms $\epsilon$-greedy in our new items recommendation scenario, to design an $\epsilon$-greedy consistent mechanism doesn't get the best social welfare. A future work would be designing a truthful mechanism that is consistent with Thompson sampling algorithm, which will greatly improve the social welfare, i.e, minimizing the regret. Also, characterizing multi-round truthful mechanisms is a promising theoretical problem to be studied.

# REFERENCES


[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[2] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88. ACM, 2009.

[3] G. Bahar, R. Smorodinsky, and M. Tennenholtz. Economic recommendation systems. *arXiv preprint arXiv:1507.07191*, 2015.

[4] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[5] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework, results and applications. In *Proceedings of the 30th international conference on machine learning*, pages 151–159, 2013.

[6] P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22. ACM, 2014.

[7] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

[8] L. Han, D. Kempe, and R. Qiang. Incentivizing exploration with heterogeneous value of money. In *International Conference on Web and Internet Economics*, pages 370–383. Springer, 2015.

[9] C.-J. Ho, A. Slivkins, and J. W. Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, 55:317–359, 2016.

[10] I. Kremer, Y. Mansour, and M. Perry. Implementing the wisdom of the crowd. *Journal of Political Economy*, 122(5):988–1012, 2014.

[11] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[12] L. Li, W. Chu, J. Langford, and R. 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. ACM, 2010.

[13] Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582. ACM, 2015.

[14] E. S. Maskin. Mechanism design: How to implement social goals. *The American Economic Review*, 98(3):567–576, 2008.

[15] R. B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

[16] R. B. Myerson. Mechanism design. In *Allocation, Information and Markets*, pages 191–206. Springer, 1989.

[17] F. Radlinski, R. Kleinberg, and T. Joachims. Learning

diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[18] H. Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

[19] Y. Shoham and K. Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[20] A. Singla and A. Krause. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1167–1178. ACM, 2013.