2021 IEEE 37th International Conference on Data Engineering (ICDE) | 978-1-7281-9184-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICDE51399.2021.00029

Crowdsensing Data Trading based on Combinatorial Multi-Armed Bandit and Stackelberg Game

Baoyi An*†, Mingjun Xiao*†, An Liu‡, Xike Xie*†, and Xiaofang Zhou§
*School of Computer Science and Technology, University of Science and Technology of China, China
†Suzhou Institute for Advanced Research, University of Science and Technology of China, China
‡Computer Science and Technology, Soochow University, China
§Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
Email: *†{anbaoyi@mail, xiaomj@, xkxie@}.ustc.edu.cn ‡anliu@suda.edu.cn §zxf@cse.ust.hk

***Abstract*—Crowdsensing Data Trading (CDT), through which a platform can aggregate some data collected by a group of mobile users with sensing devices (a.k.a., data sellers) and sell the corresponding statistics to data consumers, has been recognized as a promising paradigm for large-scale data trading in recent years. It is critical to select sellers with high sensing qualities and maximize all trading participants' profits simultaneously. However, most existing CDT systems either assume that sellers' sensing qualities are known in advance or cannot realize concurrent profit maximization. In this paper, we propose a data trading mechanism based on Combinatorial Multi-Armed Bandit and three-stage Hierarchical Stackelberg game, called CMAB-HS, to tackle the problem of quality unknown seller selection and incentive strategy design. Our objective is to select a group of sellers to maximize the total sensing quality within time budget, and determine the optimal incentive strategy for each participant to maximize individual profit simultaneously. We theoretically prove that CMAB-HS achieves Stackelberg Equilibrium and a tight bound on regret. Additionally, we demonstrate its significant performances through extensive simulations on real data traces.**

***Index Terms*—Crowdsensing data trading, Combinatorial multi-armed bandits, Stackelberg game, Online learning**

I. Introduction

Due to the research and analysis purposes, businesses and individuals have an increasing data demands and consider to purchase data from some data trading systems. However, most of the systems cannot offer appropriate data required by data consumers. To tackle the dilemma in data trading, some data trading systems (e.g., Thingful [1], Thingspeak [2]) consider to adopt Mobile CrowdSensing [3], in which a crowd of mobile users are recruited to collect location-sensitive data with their carried smart devices when they visit some pre-designated places. This data trading scheme is also called Crowdsensing Data Trading (CDT), which has greater advantages than traditional data trading for collecting data with economic value distributed in a broad-scale area by leveraging users' mobility and diverse sensing devices.

A typical CDT system consists of three parties: a platform working as the data trading broker, some data sellers, and some data consumers, as shown in Fig. 1. The platform can select some sellers to collect data from the specific Point of Interests (PoIs) assigned by consumers, where collecting data is also called sensing. Besides the data collection service, the platform can provide data aggregation service for some consumers who prefer to purchase the data statistics rather than the original chaotic data, because the consumers may not be able to analyze the massive data by themselves.

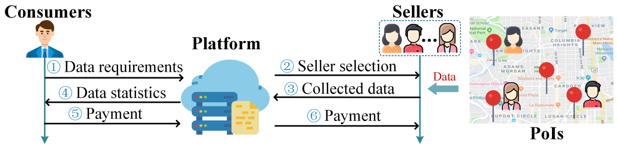


Fig. 1: The crowsensing data trading system

Since the valuation of collected data and statistics is positively correlated to the sellers' sensing qualities, it is critical to select a group of sellers with the highest qualities. However, most existing CDT systems either do not consider sellers' qualities or assume the quality information to be known in advance. For example, the CDT systems proposed in [4]–[8] select sellers based on cost with no regard of quality. [9] and [10] deem the sellers' qualities as known and unchanged information in off-line CDT systems. [11] proposes a quality-aware online CDT system which updates sellers' qualities according to the collected data, but the initial qualities are also assumed to be known. Actually, it is very challenging to acquire sellers' sensing qualities in practice, so that we need to tackle the seller selection problem with unknown qualities.

On the other hand, collecting data will inevitably incur some costs due to the usage of sensing devices, so the CDT system needs to provide adequate monetary rewards to incentivize participants to take part in the data trading. Actually, many incentive mechanisms have been proposed, especially for crowdsensing systems. These works take various design goals into consideration, such as social welfare maximization [5], [9], cost minimization [4], [12], and quality maximization [6], [11], etc. Most of these state-of-the-art works only involve two parties (*i.e.*, the platform and data sellers). Even though the works in [8], [13] consider data trading among three parties, they still separate it into two types of double-side data trading. However, a CDT system generally involves three parties, each of which might affect the other two parties. These traditional incentive mechanisms, which only deal with the trading between two parties, cannot work well in real CDT systems. Thus, it is necessary to design incentive mechanisms that can balance the profits of all participants.

978-1-7281-9184-3/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDE51399.2021.00029

In this paper, we focus on the CDT system design with the above-mentioned concerns, including two major challenges. The first is how to select a group of sellers to obtain the data collection qualities as high as possible under the circumstance that the sensing qualities of all sellers are unknown. In order to select appropriate sellers, the platform in CDT needs to divide the data collection into multiple rounds and learn the knowledge of sellers' qualities round by round on one hand (so-called exploration). On the other hand, the platform might also exploit learnt knowledge to select best ones among known sellers (so-called exploitation). This is actually an online learning and decision-making process. We need to design a mechanism to balance the exploration and exploitation well, so as to maximize the total sensing quality. The second challenge is how to design an incentive mechanism and derive the optimal incentive strategy to maximize each participant's profit simultaneously. Meanwhile, it also needs to guarantee that no one can improve its profit by deviating from the strategy.

Example: A data consumer wants to purchase the long-term image data about some PoIs distributed in a large-scale region via a CDT system (e.g., for machine learning model training, environment monitoring, etc.). Then, the CDT system recruits some sellers to collect the data by taking pictures around these PoIs within a time duration and aggregates them for the consumer. The system might not be familiar with these sellers, so that their sensing qualities are unknown. To achieve sufficient qualities, the system needs to incentivize sellers to participate in the data collection by providing some rewards.

To address the above challenges, we propose a data trading mechanism based on Combinatorial Multi-Armed Bandit (CMAB) and Hierarchical Stackelberg (HS) game, called CMAB-HS. First, we model the seller selection with unknown sensing quality as a CMAB problem by regarding each seller as a CMAB arm and the sensing quality as the corresponding revenue. The seller selection is thus formulated as the problem of determining a combinatorial arm-pulling policy. Then, we extend the classical concept of Upper Confidence Bound (UCB) to our CMAB scenario, and design a UCB-based greedy policy to solve the seller selection problem. Also, we model the problem of finding optimal incentive strategy as a three-stage HS game by regarding the consumer as the first tier leader, the platform as the second tier leader, and sellers as the followers. Through a backward deduction approach, we derive an optimal incentive strategy, which constitutes a unique Stackelberg Equilibrium (SE). Overall, the major contributions are summarized as follows:

1) We propose a data trading mechanism based on CMAB and three-stage HS game, namely CMAB-HS. To the best of our knowledge, this is the first work that combines CMAB and HS to solve the quality unknown seller selection and the optimal incentive strategy design problems in CDT systems.
2) We design a UCB-based greedy algorithm to select unknown sellers, whereby CMAB-HS can maximize the total quality revenue as much as possible. Moreover, we analyze the online performance of CMAB-HS and derive a tight upper bound on the expected regret.
3) We design an incentive mechanism based on three-stage HS game and derive an optimal incentive strategy, whereby each participant can maximize its profit. This optimal strategy constitutes a unique SE, so that no one can improve its profit by deviating from this strategy.
4) We conduct extensive simulations on real data traces to demonstrate the performance of CMAB-HS.

The remainder of the paper is organized as follows. In Sec. II, we introduce the system modeling and the problem formulation. The detailed design and theoretical analysis of CMAB-HS are elaborated in Sec. III and Sec. IV. The simulations and evaluations are presented in Sec. V. We review the related works in Sec. VI, and conclude the paper in Sec. VII.

II. System Overview, Modeling, and Problems

A. System Overview

We consider a CDT system, which is composed of a platform, some data consumers, and a crowd of unknown data sellers. A consumer can purchase data statistics from the CDT system by recruiting some sellers to collect the raw sensing data (e.g., traffic, noise, air quality data, etc) in an urban area periodically before a deadline. The consumer, the platform, and unknown sellers are defined as follows:

Definition 1 (**Consumer, Job, and Round**). The consumer is a data service requester who wants to buy the statistics on some location-sensitive data. The data can be obtained by a long-term data collection job $Job \overset{\text{def}}{=} \langle \mathcal{L}, N, T, Des \rangle$, where $\mathcal{L} = \{1, 2, \cdots, L\}$ includes L PoIs in an urban area and Des describes the requirements for collected data and aggregated statistics. The data collection job is divided into N rounds, denoted by $t \in \{1, 2, \cdots, N\}$, each of which lasts for a duration of T. That is, the whole duration of data trading is NT.

Definition 2 (**Platform**). The platform acts as a data trading broker in the CDT system. It receives the data service request from the consumer, selects some data sellers to collect raw data, aggregates these collected data, and provides the statistics to the consumer. Meanwhile, it will charge some monetary rewards from the consumer for providing the data service, from which it will extract a part of rewards in proportion as its own commission to compensate for the cost of aggregating raw data. Also, it will pay the selected sellers the remaining rewards to compensate for their data collection costs. As the broker, the platform might manipulate the payments to sellers.

Definition 3 (**Unknown Seller, Sensing Quality, and Sensing Time**). The CDT system includes M unknown data sellers, denoted by $\mathcal{M} = \{1, 2, ..., M\}$. We let $q_{i,l}^t \in [0, 1]$ denote the sensing quality of seller i $(\in \mathcal{M})$ completing data collection at PoI l $(\in \mathcal{L})$ in the t-th round. Each $q_{i,l}^t$ follows an unknown distribution with an unknown expectation q_i, and will be used to learn the estimated sensing quality of seller i in the t-th round, denoted by $\bar{q}_i^t$. We thus say that seller i is unknown. Here, we assume that the expected quality q_i only depends on seller i's smart device. And, each seller i will collect data

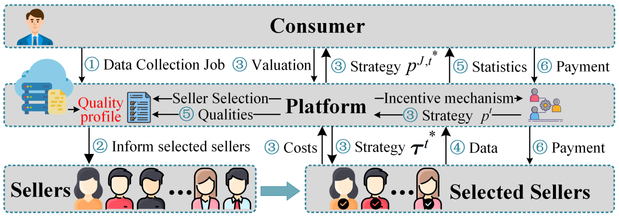


Fig. 2: The workflow of CDT in round t

from all L PoIs in each round t. The total sensing time that the seller i contributes to the data collection in round t is τ_i^t, where $\tau_i^t \in [0, T]$. Here, τ_i^t does not need to cover the entire round. Its duration can be determined by the seller itself. The longer the sensing time, the larger the size of data collected by the seller will be. Let $\boldsymbol{\tau_i}{=}(\tau_i^1, \tau_i^2, \cdots, \tau_i^N)$ be the sensing time vector of seller i in all rounds, and $\boldsymbol{\tau^t}{=}(\tau_1^t, \tau_2^t, \cdots, \tau_M^t)$ be the sensing time vector of M sellers in the t-th round.

Remark. In Def. 3, the expected quality q_i is assumed to be fixed, but the actual sensing quality $q_{i,l}^t$ might be affected by some exogenous factors (e.g., personal willingness, sensing context, daily routine, and etc.). Hence, the sensing quality $q_{i,l}^t$ might deviate from the expected quality. As the example of taking pictures in Sec. I, q_i depends on the lens, pixel, software and etc. of the camera embedded in its device, which is a fixed value. But the distance and angle of taking picture will make $q_{i,l}^t$ vary in different places even with the same device. That is, for $\forall$ task $l'{\neq}l$, $q_{i,l'}^t$ may not be equal to $q_{i,l}^t$.

The commodities for sale in CDT are the services of data aggregation and data collection from the platform and each seller, respectively. Providing these services will incur some costs for sellers and the platform. The data statistics will also produce a certain economic valuation for the consumer. The costs and valuation are defined as follows:

Definition 4 (Cost, and Valuation). Each seller $i \in \mathcal{M}$ has a cost for collecting data in round t, which can be seen as a function about the sensing time τ_i^t and learned quality $\bar{q}_i^t$, denoted as $C_i(\tau_i^t, \bar{q}_i^t)$. Likewise, the platform will incur a service cost for aggregating data, which is determined by the data size and thus is positively correlated to the sensing time of all sellers, denoted as $C^J(\boldsymbol{\tau^t})$. The consumer can obtain a valuation from the aggregated statistics, denoted as $\phi(\boldsymbol{\tau^t}, \bar{q}^t)$, which can be seen as a function of the sensing time $\boldsymbol{\tau^t}$ and the overall mean quality $\bar{q}^t$. Note that $\bar{q}^t$ is the average value of all selected sellers' learned qualities $\{\bar{q}_i^t\}_{i=1}^M$ in round t.

To encourage the consumer, platform, and sellers to participate in CDT, an incentive mechanism is designed as follows:

Definition 5 (Incentive Mechanism, Unit Service Price, and Incentive Strategy). In the CDT system, the consumer will offer a monetary reward, which is no larger than the valuation of received data statistics, to compensate for the service costs of the platform and selected sellers. Since both of the data collection and aggregation costs are related to the sensing time, the reward will be divided in proportion to the sensing time and be paid to the platform and sellers, respectively. Thus, in each round t, the consumer determines a unit data service price (*i.e.*, the reward per sensing time for data collection and aggregation), denoted as $p^{J,t}{\in}[p_{min}^J, p_{max}^J]$. The platform also determines a unit data collection service price for sellers, denoted as $p^t \in [p_{min}, p_{max}]$. Moreover, each selected seller i will determine its sensing time τ_i^t. After receiving the data statistics, the consumer will pay the platform the reward which equals to the unit data service price multiplied by the total sensing time. Also, the platform will pay each seller the reward which is the unit data collection service price multiplied by its sensing time. Since the whole incentive mechanism mainly depends on the unit service prices and sellers' sensing time, we call the triple $\langle p^{J,t}, p^t, \boldsymbol{\tau^t} \rangle$ the incentive strategy. Once the incentive strategy is determined, all payments can be settled.

We illustrate the detailed workflow of CDT in Fig.2. First, the consumer starts the data trading by publishing the data collection job to the platform, which will be conducted round by round until the given time is exhausted. Then, the platform selects some unknown sellers with high sensing qualities. Next, the consumer, the platform, and the selected sellers will cooperatively determine an incentive strategy. Meanwhile, the selected sellers go to collect data with their carried sensing devices and return the results to the platform. After that, the platform will aggregate the collected data, send the statistics to the consumer, and update the qualities based on statistics. Finally, the consumer will pay the rewards to the platform and sellers according to the incentive mechanism. Here, we list major notations in Table I for ease of reference.

Problems. There are two major challenges in the above CDT system: the *unknown seller selection* problem and the *optimal incentive strategy* problem. The first is how to select a group of unknown sellers with the highest sensing qualities. Actually, seller selection is an online learning and decision process. The platform can repeatedly learn the knowledge about sellers' sensing qualities on one hand, generally called exploration. On the other hand, it can also exploit the learnt knowledge to select the sellers with the known largest sensing qualities, generally called exploitation. We need to find the best balance between the exploration and exploitation, so as to maximize the total sensing quality as much as possible. The second is how to determine an optimal incentive strategy. Note that the consumer, the platform, and each seller can affect the others' profits and meanwhile maximize their own profits by strategically manipulate their unit prices and sensing time (*i.e.*, $p^{J,t}, p^t, \boldsymbol{\tau^t}$), respectively. There exists a game among the three parties. We need to find an optimal equilibrium for this game, so that each party can achieve its maximum profit.

B. Modeling and Formulation of Seller Selection

Selecting sellers with the largest sensing qualities under the circumstance that their qualities are unknown a priori is a critical issue in CDT. Since this is actually an online learning and decision-making process, we model it as a Combinatorial Multi-Armed Bandit (CMAB) problem. CMAB is a widely-used reinforcement learning model for online decision-making in uncertain environments [14]. It basically includes a slot machine with multiple arms, each of which is associated with

TABLE I: Description of major notations

Variable	Description
l, i, t	the indexes of PoI, seller and round, respectively
$\mathcal{L}$, $\mathcal{M}$	the sets of PoIs and sellers, respectively
L, M	the numbers of PoIs and sellers, respectively
T, N	the duration of one round, and the number of total rounds
χ_i^t	indicates whether seller i is selected in round t
K, $\mathcal{S}^t$	the number and set of selected sellers in round t
$q_{i,l}^t$	the quality of seller i collecting data at PoI l in round t
$\bar{q}_i^t, \hat{q}_i^t$	the learned quality and UCB value of seller i in round t
$\bar{q}^t$	the overall mean quality of all selected sellers in round t
n_i^t	the total learned times of seller i's quality until round t
τ_i^t	the sensing time of seller i in round t
p^t, $p^{J,t}$	the unit prices of data collection and service in round t
τ_i^{t*}	the strategy of seller i in round t
p^{t*}, $p^{J,t*}$	the strategies of platform and consumer in round t
$C_i(\cdot), a_i, b_i$	the cost function and parameters of seller i
$C^J(\cdot), \theta, \lambda$	the cost function and parameters of platform
$\phi(\cdot), \omega$	the valuation function and parameter of consumer
$\Psi_i^t(\cdot)$	the profit function of seller i in round t
$\Omega^t(\cdot)$,$\Phi^t(\cdot)$	the profit functions of platform and consumer in round t

a revenue drawn from an unknown distribution. A player will pull some arms round by round according to a bandit policy, so as to maximize the cumulative revenue. To the end, we can define the modeling of unknown seller selection.

Definition 6 (**Unknown Seller Selection Modeling**). We model the unknown seller selection as a K-armed CMAB game, where the platform is treated as the player, each seller in $\mathcal{M}$ is an arm, selecting a seller is equivalent to pulling the corresponding arm, and each seller's sensing quality is seen as the revenue of pulling the related arm. In each round, the platform selects K sellers by pulling K arms simultaneously.

After the modeling, selecting sellers with the largest qualities becomes to determine an arm-pulling policy which can maximize the total expected revenue, generally called bandit policy. The bandit policy and revenue are defined as follows:

Definition 7 (**Bandit Policy**). A bandit policy $\boldsymbol{\chi}$ is a sequence of arm-pulling decisions, which can be represented as an indicator vector $(\boldsymbol{\chi^1}, \cdots, \boldsymbol{\chi^t}, \cdots, \boldsymbol{\chi^N})$, where $\boldsymbol{\chi^t} = (\chi_1^t, \chi_2^t, \cdots, \chi_M^t) \in \{0,1\}^M$ and N is the total rounds. Moreover, $\chi_i^t = 1$ indicates that the seller i will be selected in the t-th round, while $\chi_i^t = 0$ means that it will not be selected.

Definition 8 (**Revenue**). The total revenue refers to the total sensing qualities of the sellers selected by a given bandit policy $\boldsymbol{\chi}$, denoted by $R(\boldsymbol{\chi})$. Then, the total expected revenue is:

$$E[R(\boldsymbol{\chi})] = \sum_{t=1}^{N} \sum_{i=1}^{M} \sum_{l=1}^{L} q_{i,l}^t \chi_i^t \tag{1}$$

Now, the unknown seller selection can be formulated:

$$Maximize: \quad E[R(\boldsymbol{\chi})] \tag{2}$$

$$Subject\ to: \quad \sum_{i=1}^{M} \chi_i^t = K, \forall t \in [1, N] \tag{3}$$

$$\chi_i^t \in \{0,1\}, \forall i \in \mathcal{M}, \forall t \in [1, N] \tag{4}$$

Here, Eqs. (3) and (4) indicate that K sellers are selected.

C. Modeling and Formulation of Incentive Strategy

In the CDT system, the consumer can manipulate the unit service price $p^{J,t}$ to dominate the rewards paid to the platform and sellers. The platform can also manipulate the unit data collection price p^t to determine sellers' incomes. Meanwhile, sellers can affect the profits of the consumer and platform by adjusting their own sensing time $\boldsymbol{\tau^t}$, in turn. In order to derive the optimal incentive strategy, denoted by $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau^{t*}} \rangle$, we model it as a three-stage Hierarchical Stackelberg (HS) game. First, we define the profit for each participant:

Definition 9 (**Seller's Profit**). The profit of each seller i is the difference between the payment from the platform and its data collection cost in each round t, which is defined:

$$\Psi_i^t(p^t, \tau_i^t) = p^t \tau_i^t \chi_i^t - C_i(\tau_i^t, \bar{q}_i^t) \chi_i^t. \tag{5}$$

In Eq. (5), the first part is seller i's reward, and the second part is the data collection cost. The cost function $C_i(\tau_i^t, \bar{q}_i^t)$ in the second part is assumed to be a monotonically increasing, differentiable and strictly convex function. In this paper, we adopt a widely used quadratic cost function, like in [15]–[17]:

$$C_i(\tau_i^t, \bar{q}_i^t) = (a_i {\tau_i^t}^2 + b_i \tau_i^t) \bar{q}_i^t, \tag{6}$$

$C_i(\cdot)$ reflects seller i's effort level (*i.e.*, sensing time in this paper and [15], [16], and participation level in [17]) on data collection with $a_i > 0, b_i \geq 0$. Note that $C_i(\cdot)$ increases with the sensing time and the growth rate of $C_i(\cdot)$ also increases with the sensing time, which can be used to model the increasing marginal cost for every additional unit of effort exerted. $\bar{q}_i^t \in [0, 1]$ is seller i's estimated quality currently.

Definition 10 (**Profit of Platform**). The profit of platform is the reward minus the data collection and aggregation costs:

$$\Omega^t(p^{J,t}, p^t, \boldsymbol{\tau^t}) = p^{J,t} \sum_{i=1}^{M} \tau_i^t \chi_i^t - p^t \sum_{i=1}^{M} \tau_i^t \chi_i^t - C^J(\boldsymbol{\tau^t}). \tag{7}$$

Here, the first part is the total rewards from consumer, the second part is the total payments to sellers, and the third part is the data aggregation cost. Similarly, we also adopt a quadratic function to model the aggregation cost $C^J(\boldsymbol{\tau^t})$:

$$C^J(\boldsymbol{\tau^t}) = \theta \big(\sum_{i=1}^{M} \tau_i^t \chi_i^t \big)^2 + \lambda \sum_{i=1}^{M} \tau_i^t \chi_i^t, \tag{8}$$

where $\theta > 0, \lambda \geq 0$ are pre-defined parameters.

Definition 11 (**Consumer's Profit**). The consumer's profit is the difference between the valuation of received data statistics and the rewards paid to the platform and sellers:

$$\Phi^t(p^{J,t}, \boldsymbol{\tau^t}) = \phi(\boldsymbol{\tau^t}, \bar{q}^t) - p^{J,t} \sum_{i=1}^{M} \tau_i^t \chi_i^t. \tag{9}$$

In Eq. (9), the first part is the valuation produced by data statistics, and the second part is the total payments. The valuation function $\phi(\boldsymbol{\tau^t}, \bar{q}^t)$ is assumed to be a monotonically increasing, differentiable and strictly concave function of $\boldsymbol{\tau^t} = (\tau_i^t)_{\forall i \in \mathcal{M}}$. We adopt a similar valuation function as in [16], [18]–[21]:

$$\phi(\boldsymbol{\tau^t}, \bar{q}^t) = \omega \cdot \ln \big(1 + \bar{q}^t \sum_{i=1}^{M} \tau_i^t \chi_i^t \big), \tag{10}$$

where $\omega > 1$ is a system parameter and $\bar{q}^t = \frac{\sum_{i=1}^{M} \bar{q}_i^t \chi_i^t}{\sum_{i=1}^{M} \chi_i^t}$ is the mean of estimated sensing qualities of selected sellers in the t-th round. Note that $\phi(\cdot)$ increases with the sensing time but the growth rate of $\phi(\cdot)$ decreases with the sensing time, which is known as the diminishing marginal return.

Then, we model the optimal incentive strategy as follows:

Definition 12 (**Incentive Strategy Modeling**)**.** Determining the optimal incentive strategy $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau^{t*}} \rangle$ is modeled as a three-stage Hierarchical Stackelberg (HS) game, where the consumer is the first tier leader, the platform is the second tier leader, and sellers are the followers. Each of them tries to maximize its own profit by determining an optimal parameter in the incentive strategy which can be manipulated by itself (called its *optimal strategy* hereafter, for simplicity), satisfying:

Stage 1 [Consumer's Side]: $p^{J,t^*}{=}\arg\max_{p^{J,t}} \Phi^t(p^{J,t}, \boldsymbol{\tau^t})$ (11)

Stage 2 [Platform's Side]: $p^{t^*}{=}\arg\max_{p^t} \Omega^t(p^{J,t}, p^t, \boldsymbol{\tau^t})$ (12)

Stage 3 [Seller i's Side]: $\tau_i^{t^*}{=}\arg\max_{\tau_i^t} \Psi_i^t(p^t, \tau_i^t)$ (13)

In the above game, our objective is to find an optimal incentive strategy $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau^{t*}} \rangle$, by which each participant can maximize its own profit. Meanwhile, the optimal solution must satisfy the Stackelberg Equilibrium (SE), so that no one is willing to adopt other strategies which will lead to a less profit. The SE is defined as follows:

Definition 13. (**Stackelberg Equilibrium, SE**). An optimal incentive strategy $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau^{t*}} \rangle$ constitutes a SE *iff* the following set of inequalities is satisfied:

$$\Phi^t(p^{J,t^*}, \boldsymbol{\tau^{t}}^*) \geq \Phi^t(p^{J,t}, \boldsymbol{\tau^{t}}^*), \tag{14}$$

$$\Omega^t(p^{J,t^*}, p^{t^*}, \boldsymbol{\tau^{t}}^*) \geq \Omega^t(p^{J,t^*}, p^t, \boldsymbol{\tau^{t}}^*), \tag{15}$$

$$\Psi_i^t(p^{t^*}, \tau_i^{t^*}) \geq \Psi_i^t(p^{t^*}, \tau_i^t, \boldsymbol{\tau_{-i}^{t}}^*), \tag{16}$$

where $\boldsymbol{\tau_{-i}^{t}}^*$ denotes the optimal strategies of all sellers except the seller i. Therefore, all sellers' optimal strategies can be represented as $\boldsymbol{\tau^{t}}^*{=}\boldsymbol{\tau_{-i}^{t}}^* \cup \{\tau_i^{t^*}\}$. Def. 13 shows that no one can improve its own profit by deviating from the optimal strategy.

III. The CMAB-HS Data Trading Mechanism

In this section, we propose the CMAB-HS data trading mechanism to solve the unknown seller selection problem and determine the optimal incentive strategy. First, we extend the traditional UCB mechanism from the multi-armed bandit scenario of pulling single arm to solve our K-armed CMAB problem. Next, we derive the optimal incentive strategy by using the backward induction approach. In the following, we first introduce the basic idea of CMAB-HS, and then present the detailed algorithm, followed by an illustrative example to show how our CMAB-HS mechanism works.

A. Unknown Seller Selection

An extended UCB-based bandit policy is designed to select unknown sellers for our K-armed CMAB problem. The UCB value of each seller is composed of the seller's currently estimated sensing quality and the corresponding confidence upper bound. Since the UCB value takes account of the knowledge learned from previous rounds (*i.e.*, estimated sensing quality) and the uncertainty (*i.e.*, the confidence), it can balance the exploration and exploitation well in online decision.

For an arbitrary t-th round, we estimate each seller's quality based on the knowledge learned from previous rounds. Let $\bar{q}_i^t$ be seller i's currently estimated quality and n_i^t be the number of times that seller i's quality has been learned. Then, they can be iteratively computed as the following formulations:

$$n_i^t = \begin{cases} n_i^{t-1} + L, & \chi_i^t = 1 \\ n_i^{t-1}, & \chi_i^t = 0 \end{cases} \tag{17}$$

$$\bar{q}_i^t = \begin{cases} \frac{\bar{q}_i^{t-1} n_i^{t-1} + \sum_{l \in \mathcal{L}} q_{i,l}^t}{n_i^{t-1} + L}, & \chi_{i,t} = 1 \\ \bar{q}_i^{t-1}, & \chi_{i,t} = 0 \end{cases} \tag{18}$$

Here, in Eq. (17), L indicates that once a seller i is selected in a round, its sensing quality would be learned L times. This is because the seller collects data from all of the L PoIs. Each $q_{i,l}^t$ in Eq. (18) is the corresponding sensing quality value.

Next, we compute the UCB value for each seller under the K-armed CMAB scenario as follows:

$$\hat{q}_i^t = \bar{q}_i^t + \varepsilon_i^t, \quad \varepsilon_i^t = \sqrt{\frac{(K{+}1) \cdot \ln(\sum_{j \in \mathcal{M}} n_j^t)}{n_i^t}} \tag{19}$$

Here, $\bar{q}_i^t$ is seller i's estimated quality value, and ε_i^t is the corresponding confidence upper bound. The additive factor ε_i^t takes the uncertainty of estimation into consideration, which can make the less selected sellers before have more chances to be selected in the current round. After all sellers' UCB values are calculated, we always select the K sellers with the largest UCB values in each round. In the following section, we will show that such a bandit policy can achieve the nearly optimal online decision performance.

B. Determining the Optimal Incentive Strategy

To determine the optimal incentive strategy $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau^{t*}} \rangle$, we adopt the backward induction approach to derive the solutions to Eqs. (11)-(13) in Def. 12. First, we investigate the third Stage of game to derive each seller's optimal strategy (*i.e.*, the sensing time $\tau_i^{t^*}$) for any given unit data collection service price p^t (*i.e.*, the strategy of platform). Next, we consider the second Stage of game to determine the optimal strategy of platform p^{t^*} for any given unit service price $p^{J,t}$ (*i.e.*, the consumer's strategy). Finally, we back to the first Stage of game to find the consumer's optimal strategy, *i.e.*, p^{J,t^*}. The detailed deduction is presented as follows.

Theorem 14. *In Stage 3, given any unit data collection service price p^t, each seller i's optimal strategy i can be determined:*

$$\tau_i^{t^*} = \tfrac{p^t - \bar{q}_i^t b_i}{2\bar{q}_i^t a_i}. \tag{20}$$

Proof. By deriving the first-order and second-order derivatives of each seller's profit function $\Psi_i^t(p^t, \tau_i^t)$ in Eq. (5) with respect to τ_i^t, we can derive that $\frac{\partial^2 \Psi_i^t(p^t, \tau_i^t)}{\partial (\tau_i^t)^2}{=}{-}2\bar{q}_i^t a_i{<}0$, which means that $\Psi_i^t(p^t, \tau_i^t)$ is strictly concave in the feasible region of τ_i^t. We can obtain the unique optimal strategy of each seller i by solving $\frac{\partial \Psi_i^t(p^t, \tau_i^t)}{\partial \tau_i^t}{=}0$. □

Theorem 15. *In Stage 2, based on sellers' optimal strategies determined in Stage 3, the platform can give its optimal strategy p^{t^*} for all selected sellers as follows:*

$$p^{t^*} = \tfrac{p^{J,t} A - (\lambda A - 2\theta B A + B)}{2A(1 + \theta A)}, \tag{21}$$

where $A{=}\sum_{i=1}^K \frac{1}{2\bar{q}_i^t a_i}$ and $B{=}\sum_{i=1}^K \frac{b_i}{2a_i}$.

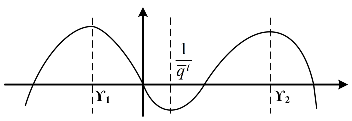


Fig. 3: The consumer's profit function $\Phi^t(\Upsilon)$

Proof. By substituting Eq. (20) into the platform's profit function in Eq. (7) and deriving the first-order and second-order derivatives of $\Omega^t(p^{J,t}, p^t, \boldsymbol{\tau^t})$ with respect to p^t, we can get $\frac{\partial^2\Omega^t(p^{J,t},p^t,\boldsymbol{\tau^t})}{\partial(p^t)^2}=-2A-2\theta A^2<0$, where $A=\sum_{i=1}^{K}\frac{1}{2\bar{q}_i^t a_i}$. Hence, $\Omega^t(p^{J,t}, p^t, \boldsymbol{\tau^t})$ is strictly concave in the feasible region of p^t. We can obtain the unique optimal strategy of platform by solving $\frac{\partial\Omega^t(p^{J,t},p^t,\boldsymbol{\tau^t})}{\partial p^t}=0$. □

Theorem 16. *In Stage 1, the consumer's optimal strategy is:*

$$p^{J,t^*}=\frac{3\bar{q}^t\Lambda+\sqrt{(\bar{q}^t\Lambda-2)^2+8\Theta\omega(\bar{q}^t)^2}-2}{4\bar{q}^t\Theta}, \tag{22}$$

where $\Theta=\frac{A}{2(1+\theta A)}$, $\Lambda=\frac{\lambda A-2\theta BA+B}{2(1+\theta A)}+B$, $A=\sum_{i=1}^{K}\frac{1}{2\bar{q}_i^t a_i}$, $B=\sum_{i=1}^{K}\frac{b_i}{2a_i}$. *If* $p^{J,t^*}>p^J_{max}, p^{J,t^*}=p^J_{max}$; $p^{J,t^*}<p^J_{min}, p^{J,t^*}=p^J_{min}$.

Proof. First, by substituting Eq. (20) into the consumer's profit function in Eq. (9), we can obtain the profit of the consumer:

$$\Phi^t(p^{J,t}, \boldsymbol{\tau^t})=\omega\ln(1-\bar{q}^tB+p_i^t\bar{q}^tA)-p^{J,t}(p_i^tA-B). \tag{23}$$

Then, by substituting Eq. (21) into Eq. (23), we can obtain:

$$\Phi^t(p^{J,t}, \boldsymbol{\tau^t}) = \omega\ln\big(1-\bar{q}^tB+\tfrac{p^{J,t}A-(\lambda A-2\theta BA+B)}{2A(1+\theta A)}\bar{q}^tA\big) \\ -p^{J,t}\big(\tfrac{p^{J,t}A-(\lambda A-2\theta BA+B)}{2A(1+\theta A)}A-B\big). \tag{24}$$

Let $\Theta=\frac{A}{2(1+\theta A)}$, $\Lambda=\frac{(\lambda A-2\theta BA+B)}{2(1+\theta A)}+B$. Eq. (24) is rewritten:

$$\Phi^t(p^{J,t}, \boldsymbol{\tau^t})=\omega\ln(1+\bar{q}^t\Theta p^{J,t}-\bar{q}^t\Lambda)-\Theta(p^{J,t})^2+\Lambda p^{J,t}. \tag{25}$$

Let $\Upsilon=\Lambda-\Theta p^{J,t}<0$. We can observe that $-\Upsilon=\sum_{i=1}^{K}\tau_i^t$, *i.e.*, the total sensing time contributed by selected sellers. Next, we convert the optimal p^{J,t^*} determination problem to optimal Υ^* determination problem, which can maximize $\Phi^t(\Upsilon)$.

$$\Phi^t(\Upsilon) = \omega\ln(1-\bar{q}^t\Upsilon)+\frac{\Upsilon(\Lambda-\Upsilon)}{\Theta}. \tag{26}$$

We derive the first-order derivative of $\Psi_i^t(\Upsilon)$ as follows:

$$\frac{\partial\Phi^t(\Upsilon)}{\partial\Upsilon}=\frac{-\omega\bar{q}^t}{1-\bar{q}^t\Upsilon}+\frac{\Lambda-2\Upsilon}{\Theta}=\frac{(\Lambda-2\Upsilon)(1-\bar{q}^t\Upsilon)-\omega\bar{q}^t\Theta}{\Theta(1-\bar{q}^t\Upsilon)} \\ =\frac{2\bar{q}^t\Upsilon^2-(\bar{q}^t\Lambda+2)\Upsilon+(\Lambda-\Theta\omega\bar{q}^t)}{\Theta(1-\bar{q}^t\Upsilon)}. \tag{27}$$

The consumer's profit function $\Phi^t(\Upsilon)$ is not standardly concave. So, to find the optimal value of Υ which maximizes $\Phi^t(\Upsilon)$, we need to analyze the monotonicity of $\Phi^t(\Upsilon)$. Let the numerator term of $\Phi^t(\Upsilon)$ be zero, *i.e.*, $2\bar{q}^t\Upsilon^2-(\bar{q}^t\Lambda+2)\Upsilon+(\Lambda-\Theta\omega\bar{q}^t)=0$. Then, we leverage the formula method to seek for the roots of this quadratic formula as follows:

$$\Delta=(\bar{q}^t\Lambda+2)^2-8\bar{q}^t(\Lambda-\Theta\omega\bar{q}^t)>(\bar{q}^t\Lambda-2)^2=0. \tag{28}$$

Since $\Delta>0$, the numerator term has two roots. When $\Lambda\geq 2/\bar{q}^t$, the two roots can be computed:

$$\Upsilon_1=\big(\bar{q}^t\Lambda+2-\sqrt{\Delta}\big)/4\bar{q}^t<\big(\bar{q}^t\Lambda+2-(\bar{q}^t\Lambda-2)\big)/4\bar{q}^t=1/\bar{q}^t, \tag{29}$$

$$\Upsilon_2=\big(\bar{q}^t\Lambda+2+\sqrt{\Delta}\big)/4\bar{q}^t>\big(\bar{q}^t\Lambda+2+(\bar{q}^t\Lambda-2)\big)/4\bar{q}^t\geq 1/\bar{q}^t. \tag{30}$$

When $\Lambda<2/\bar{q}^t$, the two roots can be computed:

$$\Upsilon_1=\big(\bar{q}^t\Lambda+2-\sqrt{\Delta}\big)/4\bar{q}^t<\big(\bar{q}^t\Lambda+2-(2-\bar{q}^t\Lambda)\big)/4\bar{q}^t<1/\bar{q}^t, \tag{31}$$

$$\Upsilon_2=\big(\bar{q}^t\Lambda+2+\sqrt{\Delta}\big)/4\bar{q}^t>\big(\bar{q}^t\Lambda+2+(2-\bar{q}^t\Lambda)\big)/4\bar{q}^t=1/\bar{q}^t. \tag{32}$$

Algorithm 1: The CMAB-HS Mechanism

```
    Input: $\mathcal{L}, \mathcal{M}, N, K, (K+1), \boldsymbol{a}, \boldsymbol{b}, \theta, \lambda, \omega, \upsilon,$
           $[p^J_{min}, p^J_{max}], [p_{min}, p_{max}], \tau^0$
    Output: $\boldsymbol{\chi}, (\boldsymbol{p^{J^*}}, \boldsymbol{p}^*, \boldsymbol{\tau}^*)$
 1  Initialize $\chi_i^t=0, p^{J,t}=0, p^t=0, \tau_i^t=0, \forall i\in\mathcal{M}, \forall t\in[1,N]$;
 2  $t=1$, select all sellers in $\mathcal{M}$ in the first round:
 3  foreach $i\in\mathcal{M}$ do  $\tau_i^{t^*}=\tau^0, \chi_i^t=1$;
 4  $p^{t^*}=p_{max}, p^{J,t^*}=\arg\min_{p^{J,t}}\Omega^t\geq 0$;
 5  foreach $i\in\mathcal{M}$ do  Update $n_i^t$, $\bar{q}_i^t$, $\hat{q}_i^t$;
 6  while $t<N$ do
 7      Sort the sellers $\mathcal{M}$ by UCB values: $\hat{q}^t_{s_1}\geq\cdots\geq\hat{q}^t_{s_M}$;
 8      $t=t+1$;
 9      Select the top $K$ sellers: $\mathcal{S}^t=\{s_1, s_2, \cdots, s_K\}$;
10      foreach $s_i\in\mathcal{S}^t$ do  $\chi^t_{s_i}=1$;
11      Execute the HS game to determine the optimal
        strategies according to Eqs. (22),(21),(20):
        $\langle p^{J,t^*}, p^{t^*}, \boldsymbol{\tau^{t*}}\rangle$;
12      foreach $i\in\mathcal{M}$ do  Update $n^t_{s_i}$, $\bar{q}^t_{s_i}$, $\hat{q}^t_{s_i}$;
13  return $(\boldsymbol{\chi}, \langle\boldsymbol{p^{J^*}}, \boldsymbol{p}^*, \boldsymbol{\tau}^*\rangle)$;
```

According to Eqs. (29)-(32), $\Upsilon_1<1/\bar{q}^t<\Upsilon_2$ always holds. The positive and negative properties of the derivative function $\frac{\partial\Phi^t(\Upsilon)}{\partial\Upsilon}$ are same with $(\Upsilon-\Upsilon_1)(1/\bar{q}^t-\Upsilon)(\Upsilon-\Upsilon_2)$. That is, $\frac{\partial\Phi^t(\Upsilon)}{\partial\Upsilon}>0$, when $\Upsilon\in(-\infty,\Upsilon_1)\bigcup(1/\bar{q}^t,\Upsilon_2)$. Otherwise, $\frac{\partial\Phi^t(\Upsilon)}{\partial\Upsilon}<0$ if $\Upsilon\in(\Upsilon_1,1/\bar{q}^t)\bigcup(\Upsilon_2,+\infty)$. Hence, the consumer's profit function $\Phi^t(\Upsilon)$ monotonically increases in $(-\infty,\Upsilon_1)$, decreases in $(\Upsilon_1,1/\bar{q}^t)$, increases in $(1/\bar{q}^t,\Upsilon_2)$, and decreases in $(\Upsilon_2,+\infty)$, as illustrated in Fig. 3. Since $\Upsilon<0$, the optimal point of $\Phi^t(\Upsilon)$ is Υ_1. So, the consumer's optimal strategy is

$$p^{J,t^*}=\frac{\Lambda-\Upsilon_1}{\Theta}=\frac{3\bar{q}^t\Lambda+\sqrt{\Delta}-2}{4\bar{q}^t\Theta}. \tag{33}$$

Additionally, according to Def. 5, $p^{J,t}\in[p^J_{min},p^J_{max}]$. Then, if $p^{J,t^*}>p^J_{max}$, we can let $p^{J,t^*}=p^J_{max}$; otherwise, if $p^{J,t^*}<p^J_{min}$, let $p^{J,t^*}=p^J_{min}$. □

According to Eq. (22), we can derive the consumer's optimal strategy p^{J,t^*}. After that, we can determine the optimal strategy p^{t^*} for the platform by substituting p^{J,t^*} into Eq. (21). Next, we can compute each seller's optimal strategy τ_i^{t*} by substituting p^{t^*} into Eq. (20). Then, the whole optimal incentive strategy is determined, based on which the consumer can pay the rewards to the platform and sellers.

C. The Detailed Algorithm

According to the above solution, the proposed CMAB-HS mechanism is depicted in Algorithm 1. At the beginning, we initialize all $\chi_i^t=0, p^{J,t}=0, p^t=0, \tau_i^t=0$ (Step 1). Then, the initial exploration phase begins. We first tentatively select all sellers to collect data in order to learn and estimate their quality values (Steps 2-4). At the end of the first round, we observe $\{q^t_{i,l}|l\in\mathcal{L}\}$ for each selected seller i, and then update $\boldsymbol{n^t}=(n_i^t)_{\forall i\in\mathcal{M}}, \boldsymbol{\bar{q}^t}=(\bar{q}_i^t)_{\forall i\in\mathcal{M}}, \boldsymbol{\hat{q}^t}=(\hat{q}_i^t)_{\forall i\in\mathcal{M}}$ according to Eqs. (17)-(19), respectively (Step 5).

	Consumer	Platform	Seller 1	Seller 2	Seller 3
Expected quality			0.8	0.7	0.6
Parameter	ω = 10	θ = 0.5, λ = 1	a_1 = 0.5, b_1 = 1	a_2 = 0.3, b_2 = 1	a_3 = 0.2, b_3 = 1

Fig. 4: System parameters and sellers' information

$q_{i,l}^t$	Round 1			Round 2		Round 3		Round 4		Round 5		Round 6		Round 7		Round 8		Round 9		Round 10	
Seller PoI	1	2	3	2	1	3	2	1	3	2	1	3	2	2	1	3	2	1	2	3	1
1	0.804	0.845	0.345	0.732	0.516	0.613	0.843	0.477	0.612	0.751	0.492	0.517	0.609	0.844	0.762	0.483	0.846	0.932	0.715	0.782	0.468
2	0.661	0.797	0.614	0.946	0.619	0.55	0.523	0.378	0.637	0.869	0.712	0.71	0.595	0.843	0.253	0.68	0.505	0.752	0.565	0.841	0.938
3	0.723	0.509	0.48	0.876	0.423	0.356	0.608	0.555	0.521	0.74	0.596	0.437	0.405	0.576	0.589	0.615	0.736	0.261	0.345	0.611	0.369
4	0.389	0.468	0.841	0.418	0.644	0.531	0.42	0.769	0.783	0.895	0.493	0.466	0.881	0.803	0.939	0.406	0.78	0.572	0.793	0.917	0.809

Fig. 5: The actual qualities in different rounds

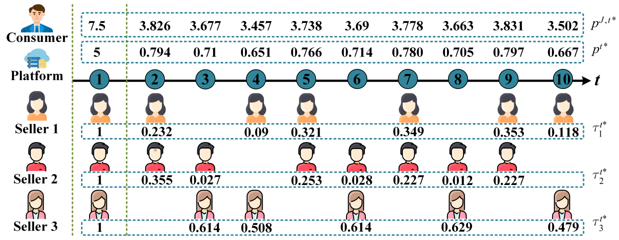


Fig. 6: The whole data trading process

After the initial exploration, we can use the learned quality information to select sellers and determine the optimal strategies of all participants (*i.e.*, exploitation). In Steps 7-10, we first sort the sellers in a non-increasing order of their UCB values, and then greedily select the top-K sellers into the selected sellers set $\mathcal{S}^t$. Next, the CMAB-HS mechanism will play the HS game among the consumer, the platform, and the selected sellers in $\mathcal{S}^t$ to determine the optimal strategies for them according to Eqs. (22), (21), and (20) (Step 11).

Besides, CMAB-HS will also update $\boldsymbol{n^t}, \boldsymbol{\bar{q}^t}, \boldsymbol{\hat{q}^t}$ simultaneously based on the observed qualities in the same current round (Step 12, *i.e.*, exploration). The CMAB-HS mechanism alternates the exploitation with the exploration in each of the subsequent rounds t ($t \in [2, N]$). So far, CMAB-HS returns the seller selection result and the strategy profile $\langle \boldsymbol{p^{J*}}{=}(p^{J,t*})_{t=1}^N, \boldsymbol{p^*}{=}(p^{t*})_{t=1}^N, \boldsymbol{\tau^*}{=}(\boldsymbol{\tau^{t*}})_{t=1}^N \rangle$.

D. An Illustrative Example

For better understanding, we provide an example to illustrate the data trading process of CMAB-HS. There are three sellers $\mathcal{M}{=}\{1,2,3\}$ and four PoIs $\mathcal{L}{=}\{1,2,3,4\}$ in a 10-rounds data trading, where $K{=}2$ sellers are selected to collect data from the four locations in each round. The system parameters and the expected qualities (which are unknown a priori) are shown in Fig. 4. We assume that the observed quality of each seller follows the Gaussian distribution in [0,1].

At the beginning, we do not have any information about sellers' qualities. Thus, according to Algorithm 1, all sellers $\langle 1,2,3 \rangle$ will be selected in the initial exploration phase to learn the quality information. Each selected seller needs to contribute one unit of time to collect data and is paid by the highest price $p^{1*} = 5$. The platform will be paid by a price $p^{J,1*}$=7.5 which can ensure its profit is non-negative. After the initial round, the sample quality means for three sellers are calculated according to Fig. 5: $\bar{q}_1^1$=(0.804+0.661+0.723+0.389)/4= $0.644, \bar{q}_2^1 = 0.654, \bar{q}_3^1 = 0.57$. Accordingly, the UCB value for each seller is calculated as: $\hat{q}_1^1 = 3.258, \hat{q}_2^1 = 3.268, \hat{q}_3^1 = 3.184$. So in round 2, sellers $\langle 2,1 \rangle$ will be selected and the corresponding strategies will be determined based on HS game, *i.e.*, $p^{J,2*}$=3.826, p^{2*}=0.794 and τ_2^{2*}=0.355, τ_1^{2*}=0.232. Then, the sellers' sample quality means and UCB values will be updated: $\bar{q}_1^2$=0.597, $\bar{q}_2^2$=0.698, $\bar{q}_3^2$=0.57, and $\hat{q}_1^2$=1.657, $\hat{q}_2^2$=1.758, $\hat{q}_3^2$=2.069. Then in the next round 3, sellers $\langle 3,2 \rangle$ will be selected, and their strategies are similarly determined. During the whole data trading, sellers will be selected in the order of $\langle 1,2,3 \rangle$, $\langle 2,1 \rangle$, $\langle 3,2 \rangle$, $\langle 1,3 \rangle$, $\langle 2,1 \rangle$, $\langle 3,2 \rangle$, $\langle 2,1 \rangle$, $\langle 3,2 \rangle$, $\langle 1,2 \rangle$, $\langle 3,1 \rangle$. We illustrate the whole data trading process in Fig. 6.

IV. Performance Analysis

In this section, we analyze the online performance of the CMAB-HS mechanism and the Stackelberg equilibrium.

A. Online Performance Analysis

First, we analyze the regret performance, which is the difference of total revenue achieved by the optimal policy and our CMAB-HS mechanism [22]. In each round t, we use the expected quality as the seller selection criteria, *i.e.*, q_i. Let $\mathcal{S}^*$ and $\mathcal{S}^t$ be the selected sellers set in optimal solution (*i.e.*, the expected qualities of sellers are known in advance) and the selected sellers set in CMAB-HS, respectively. We consider that $q_{s_1^*} \cdots \geq q_{s_K^*} \geq \cdots \geq q_{s_M^*}$, so $\mathcal{S}^* {=} \{s_1^*, s_2^* \cdots, s_K^*\}$ is always the optimal selected sellers set in each round. Here, $*$ denotes the optimal policy. Hence, the regret can be defined:

$$Reg {=} \textstyle\sum_{t=1}^N R(\boldsymbol{\chi^*}) {-} \sum_{t=1}^N E[R(\boldsymbol{\chi^t})]. \tag{34}$$

Then, we define the smallest and largest possible difference of revenue among all non-optimal selected sellers sets $\mathcal{S}^t {\neq} \mathcal{S}^*$:

$$\triangle_{max} {=} \textstyle\sum_{i \in \mathcal{S}^*} q_i - \min_{\mathcal{S}^t \neq \mathcal{S}^*} \sum_{i \in \mathcal{S}^t} q_i, \tag{35}$$

$$\triangle_{min} {=} \textstyle\sum_{i \in \mathcal{S}^*} q_i - \max_{\mathcal{S}^t \neq \mathcal{S}^*} \sum_{i \in \mathcal{S}^t} q_i. \tag{36}$$

Let β_i^t be the counter of seller i after the initial exploration (*i.e.*, $t{>}1$) . β_i^t denotes the times that seller i's quality has been learned. So $\beta_i^1 {=} L, \forall i {\in} \mathcal{M}$, and $\sum_{i \in \mathcal{M}} \beta_i^1 {=} ML$. In each round t ($t{>}1$), when $\mathcal{S}^t {\neq} \mathcal{S}^*$, the counter β_i^t is updated as follows:

$$i = \arg\min\nolimits_{j \in \mathcal{S}^t} \beta_j^{(t-1)}, \qquad \beta_i^t = \beta_i^{(t-1)} + L. \tag{37}$$

That is, we find the smallest counter of all selected sellers at the end of (t–1)-th round. If multiple sellers are satisfied the condition, we randomly select any one. So the seller with the smallest counter β_i^t will be incremented by L. This means that for $\forall i {\in} \mathcal{M}$, the sum of the counter β_i^t equals to the total quality learning times. When any non-optimal sellers set is determined in a round, there is exactly one seller's counter to be incremented. Next, we will focus on the upper bound of the counter β_i^N, where N is the total rounds of the CMAB-HS mechanism. More specifically, we have the following lemma.

Lemma 17 (Chernoff-Hoeffding bound)**.** *[14] Suppose that $X_1, X_2, \cdots, X_n$ are n random variables with common range* $[0, 1]$*, satisfying* $E\left[X_t | X_1, \cdots, X_{t-1}\right] {=} \mu$ *for* $\forall t {\in} [1, n]$*. Let* $S_n {=}$ $X_1 {+} \cdots {+} X_n$*. Then* $\forall a {\geq} 0$,

$$P\left[S_n {\geq} n\mu {+} a\right] \leq e^{-2a^2/n}, P\left[S_n {\leq} n\mu {-} a\right] \leq e^{-2a^2/n}. \tag{38}$$

Lemma 18. *The expected counter β_i^N has an upper bound for any seller $i \in \mathcal{M}$ in time rounds N, that is*

$$E[\beta_i^N] \leq \tfrac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2} + 1 + \tfrac{\pi^2}{3K^{2K+1}L^{K+2}}. \tag{39}$$

Proof. In each round t, one of the following cases must happen: 1) the optimal set of sellers, *i.e.*, $\mathcal{S}^*$, might be selected; 2) a non-optimal set will be selected, *i.e.*, $\mathcal{S}^t \neq \mathcal{S}^*$. In the first case, the counter β_i^t will not change, while in the second case, the counter β_i^t will be updated according to Eq. (37). Recall that we use $\chi_i^t \in \{0,1\}$ to denote the selection policy for seller i in the t-th round. That is, χ_i^t represents the change of the counter β_i^t, where $\chi_i^t = 1$ means that β_i^t is incremented, and χ_i^t=0 otherwise. Based on this, we have the following results:

$$\beta_i^N = L + L\sum_{t=2}^N \boldsymbol{\chi}\{\chi_i^t=1\} \leq \ell + L\sum_{t=2}^N \boldsymbol{\chi}\{\chi_i^t=1, \beta_i^t \geq \ell\} \quad (40)$$

$$\leq \ell + L\sum_{t=2}^N \boldsymbol{\chi}\Big\{ \sum_{i\in\mathcal{S}^t} \hat{q}_i^{(t-1)} \geq \sum_{i\in\mathcal{S}^*} \hat{q}_i^{(t-1)}, \beta_i^t \geq \ell \Big\} \quad (41)$$

$$\leq \ell + L\sum_{t=2}^N \boldsymbol{\chi}\Big\{ \max_{\ell \leq n_{s_1}^t \leq \cdots \leq n_{s_K}^t \leq (t-1)L} \hat{q}_{s_j}^{(t-1)} \geq \min_{L \leq n_{s_1}^* \leq \cdots \leq n_{s_K}^* \leq (t-1)L} \hat{q}_{s_j^*}^{(t-1)} \Big\} \quad (42)$$

$$\leq \ell + \sum_{t=2}^{N} \sum_{n_{s_1}^t=\ell}^{(t-1)L} \cdots \sum_{n_{s_K}^t=\ell}^{(t-1)L} \sum_{n_{s_1^*}^t=L}^{(t-1)L} \cdots \sum_{n_{s_K^*}^t=L}^{(t-1)L} \boldsymbol{\chi}\Big\{ \sum_{j=1}^K \hat{q}_{s_j}^t \geq \sum_{j=1}^K \hat{q}_{s_j^*}^t \Big\}, \quad (43)$$

where n_i^t is the number of total times that seller i's quality has been learned at the end of t-th round. According to Eq. (17), we have $n_i^t \geq \beta_i^t$, for $\forall i \in \mathcal{M}, t \in [1, N]$. Then, we focus on the bound of $\sum_{j=1}^K \hat{q}_{s_j}^t \geq \sum_{j=1}^K \hat{q}_{s_j^*}^t$. According to Eq. (19):

$$\sum_{j=1}^K \bar{q}_{s_j}^t + \varepsilon_{s_j}^t \geq \sum_{j=1}^K \bar{q}_{s_j^*}^t + \varepsilon_{s_j^*}^t. \quad (44)$$

We can obtain that at least one of the following cases must be true (which is based on the proof by contradiction):

$$\sum_{j=1}^K \bar{q}_{s_j^*}^t \leq \sum_{j=1}^K q_{s_j^*} - \varepsilon_{s_j^*}^t, \quad (45)$$

$$\sum_{j=1}^K \bar{q}_{s_j}^t \geq \sum_{j=1}^K q_{s_j} + \varepsilon_{s_j}^t, \quad (46)$$

$$\sum_{j=1}^K q_{s_j^*} < \sum_{j=1}^K q_{s_j} + 2\varepsilon_{s_j}^t. \quad (47)$$

Next, we need to prove the upper bound of the probability of Eqs. (45) and (46). By applying the Chernoff-Hoeffding bound in Lemma 17, we can get

$$\mathbb{P}\{\sum_{j=1}^K \bar{q}_{s_j^*}^t \leq \sum_{j=1}^K q_{s_j^*} - \varepsilon_{s_j^*}^t\} \leq \sum_{j=1}^K \mathbb{P}\{\bar{q}_{s_j^*}^t \leq q_{s_j^*} - \varepsilon_{s_j^*}^t\}$$

$$\leq \sum_{j=1}^K e^{-2n_{s_j^*}^t {\varepsilon_{s_j^*}^t}^2} \leq K(tKL)^{-2(K+1)}. \quad (48)$$

We can also similarly prove that

$$\mathbb{P}\{\sum_{j=1}^K \bar{q}_{s_j}^t \geq \sum_{j=1}^K q_{s_j} + \varepsilon_{s_j}^t\} \leq K(tKL)^{-2(K+1)}. \quad (49)$$

Note that, at the end of t-th round, the number of total times that all sellers' qualities have been learned is $\sum_{i\in\mathcal{M}} n_i^t = tKL$. Then, we choose a certain value ℓ to make the Eq. (47) impossible. Based on the fact that $n_i^t \geq \beta_i^t \geq \ell$, we have

$$\sum_{j=1}^K q_{s_j^*} - \sum_{j=1}^K q_{s_j} - 2\sum_{j=1}^K \varepsilon_{s_j}^t$$

$$\geq \triangle_{min} - 2\sum_{j=1}^K \sqrt{\frac{(K+1)\ln(tKL)}{n_{s_j}^t}}$$

$$\geq \triangle_{min} - 2\sum_{j=1}^K \sqrt{\frac{(K+1)\ln(tKL)}{\ell}} \geq 0. \quad (50)$$

After analyzing Eq. (50), we can yield that Eq. (50) always holds if ℓ satisfies the following condition:

$$\ell > \frac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2}. \quad (51)$$

Now we continue Eq. (43), and get

$$E[\beta_i^N] \leq \Big\lceil \frac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2} \Big\rceil + \sum_{t=1}^{+\infty} (tL-\ell)^K ((t-1)L)^K 2K(tKL)^{-2(K+1)}$$

$$\leq \frac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2} + 1 + \frac{2}{K^{2K+1}L^{K+2}} \sum_{t=1}^{+\infty} t^{-2}$$

$$\leq \frac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2} + 1 + \frac{\pi^2}{3K^{2K+1}L^{K+2}}. \quad (52)$$

Hence, the lemma holds. □

Theorem 19. *The expected regret of the CMAB-HS mechanism is bounded by* $O\big(MK^3\ln(NKL)\big)$.

Proof. According to the definition of regret in Eq. (34) and Lemma 18, we have the following result:

$$Reg = NR(\boldsymbol{\chi}^*) - E[R(\boldsymbol{\chi})] \leq \sum_{i=1}^M \beta_i^N \triangle_{max}$$

$$\leq M \triangle_{max} \left(\frac{4K^2(K+1)\ln(NKL)}{\triangle_{min}^2} + 1 + \frac{\pi^2}{3K^{2K+1}L^{K+2}} \right)$$

$$= O\big(MK^3\ln(NKL)\big). \quad (53)$$

Therefore, the theorem holds. □

B. The Stackelberg Equilibrium Analysis

Here, we prove that the existence and uniqueness of Stackelberg Equilibrium can be guaranteed in CMAB-HS.

Theorem 20. *The optimal incentive strategy* $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau}^{t*} \rangle$ *determined by the CMAB-HS mechanism constitutes the unique Stackelberg Equilibrium.*

Proof. In each round t, according to Theorem 16, the optimal strategy p^{J,t^*} of the consumer can be uniquely obtained in two cases: Case 1: $p^{J,t^*} \in [p_{min}^J, p_{max}^J]$. After p^{J,t^*} is ascertained, the optimal strategy values p^{t^*} and $\tau_i^{t^*}$ of platform and each seller i can be determined according to Eqs. (21) and (20). Since p^{J,t^*} is only calculated based on the input values of sellers' qualities $\bar{\boldsymbol{q}}^t$, cost parameters $\boldsymbol{a}, \boldsymbol{b}, \theta, \lambda$ and valuation parameter ω, the values of $\langle p^{J,t*}, p^{t*}, \boldsymbol{\tau}^{t*} \rangle$ are only associated to the constant inputs in round t. When the platform and sellers hold the optimal strategies, the consumer's profit $\Phi^t(p^{J,t}, \boldsymbol{\tau}^{t*})$ only changes with $p^{J,t}$. According to Theorem 16, the maximum profit can be obtained only at p^{J,t^*}. Any other value of $p^{J,t} \neq p^{J,t^*}$ can yield an inferior profit, so Eq. (14) holds. Similarly, when fixing p^{J,t^*} and $\boldsymbol{\tau}^{t*}$, the platform's optimal profit can be obtained only at p^{t*}; when fixing p^{J,t^*}, p^{t^*} and $\boldsymbol{\tau}_{-i}^{t*}$, the profit of seller i is maximized at τ_i^{t*}. Thus, Eqs. (15) and (16) hold. Case 2: $p^{J,t^*} \notin [p_{min}^J, p_{max}^J]$. Then $p^{J,t^*} = \{p_{min}^J, p_{max}^J\}$, which is also the fixed value. We can similarly derive that Eqs. (14)-(16) hold. Therefore, the theorem holds. □

V. Implementation and Evaluations

In this section, we evaluate the performance of CMAB-HS with extensive simulations on a real-world data trace.

A. Evaluation Methodology

Simulation Settings: We conduct extensive simulations on a real data trace of Chicago Taxi Trips [23]. Each entry of the trace records the taxiID, timestamp, trip miles and the location of picking up/dropping off passengers, etc. We choose a data

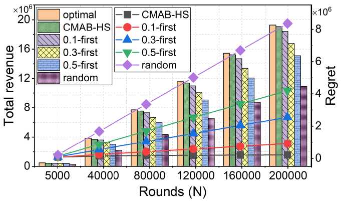
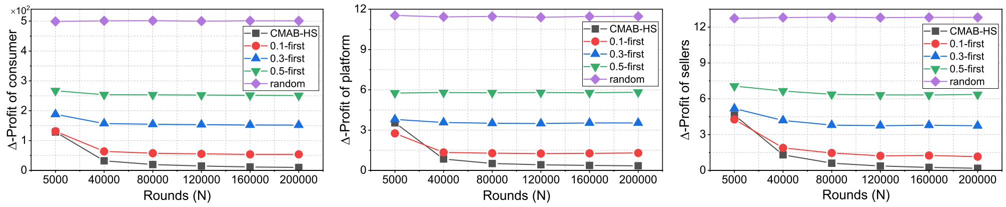


Fig. 7: Revenue and regret vs. N and $M = 300$



(a) Consumer (b) Platform (c) Sellers

Fig. 8: Δ-Profit vs. rounds N and number of sellers $M = 300$

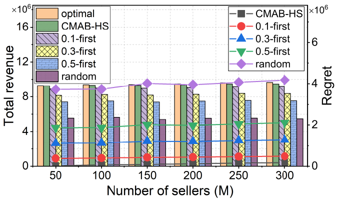
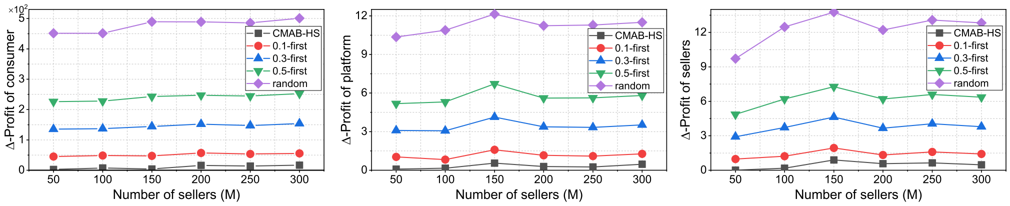


Fig. 9: Regret vs. M and $N = 10^5$



(a) Consumer (b) Platform (c) Sellers

Fig. 10: Δ-Profit vs. number of sellers M and rounds N=10^5

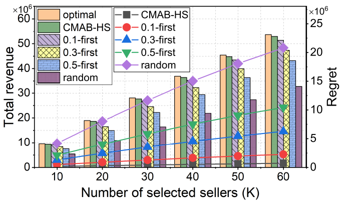
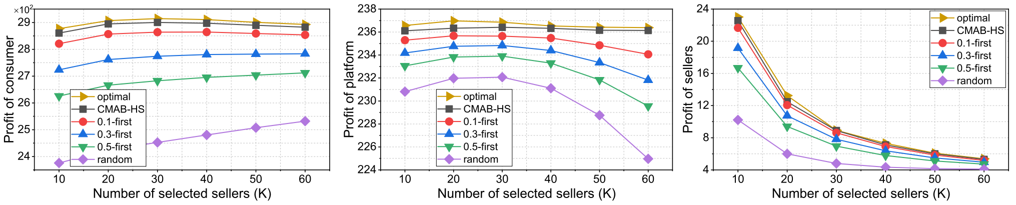


Fig. 11: Revenue and regret vs. K and $M = 300$



(a) Consumer (b) Platform (c) Sellers

Fig. 12: Profit vs. number of selected sellers K and $M = 300$

set including 27465 taxi records. In our simulation, we select some pick-up/drop-off points as the PoIs. And we assume that the taxis which pick up or drop off passengers at these points can complete the data collection job, which are regarded as the data sellers. We first choose L=10 locations and find 300 taxis from the trace. Then, we choose M taxis as satisfied sellers, where M is produced from $[50, 300]$. The number of selected sellers K is selected from $[10, 60]$. The total rounds N of online data collection job is set in $[5{\times}10^3, 2{\times}10^5]$. The default values are $M{=}300$, $K{=}10$ and $N{=}10^5$. Since there is no record about the qualities, we randomly generate the expected quality from $[0, 1]$ and then adopt truncated Gaussian distribution to generate sellers' observed qualities. Each seller i's cost is relied on its cost function parameters a_i, b_i, which are set in $[0.1, 0.5]$ and $[0.1, 1]$, respectively. Similarly, we set the cost function parameters of platform as $\theta{\in}[0.1, 1]$, $\lambda{\in}[0.5, 2]$ and the valuation function parameter of consumer as $\omega{\in}[600, 1400]$. And we set θ=0.1, λ=1, ω=1000 by default.

Compared Algorithms: There are multiple simultaneous optimization goals in our unknown online CDT scenario, while most of the existing CDT systems realize one optimization, which cannot be directly used to compare with our mechanism. Hence, we design three algorithms for comparison, called "optimal", "ϵ-first" [24], [25] and "random" [22], [25]–[27]. "optimal" means that the algorithm knows the expected qualities of all sellers in advance, and always selects the same top-K sellers with the highest qualities in each round of the data collection. "ϵ-first" will randomly select K sellers in each of the first ϵN rounds (*i.e.*, pure exploration phase) and greedily select the top-K sellers with the highest qualities in each of the remaining $(1-\epsilon)N$ rounds, where we change ϵ from 0.1 to 0.5. The random algorithm that does not know the expected qualities will randomly select K sellers in all rounds.

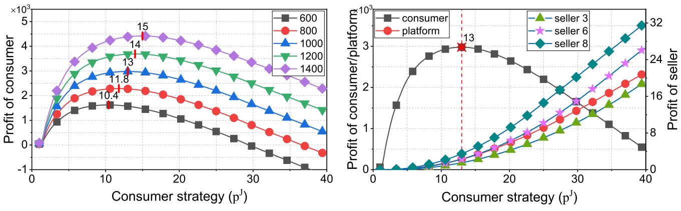
TABLE II: Simulation settings

Parameter name	Values
number of rounds N	5, 40, 80, **100**, 120, 160, 200 ($\times 10^3$)
number of sellers M	50, 100, 150, 200, 250, **300**
number of selected sellers K	**10**, 20, 30, 40, 50, 60
valuation parameter ω	600, 800, **1000**, 1200, 1400
cost parameter θ, λ	[0.1, 1], [0.5, 2]
cost parameters $\boldsymbol{a}, \boldsymbol{b}$	[0.1, 0.5], [0.1, 1]

B. Evaluation Results

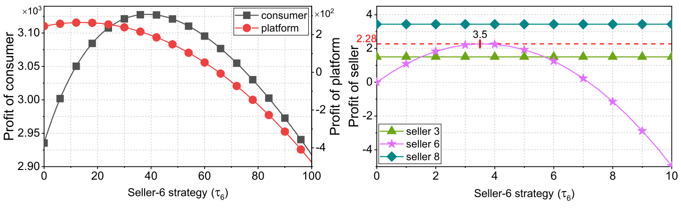
For the evaluation criteria, we adopt four main metrics: total revenue, regret, profit and strategy. Moreover, we use PoC, PoP, PoS(s), SoC, SoP and SoS(s) to denote the Profit of Consumer, Profit of Platform, Profit of selected Seller(s), Strategy of Consumer, Strategy of Platform and Strategy of selected Seller(s), respectively. For better comparison, we also define some metrics to measure the difference of profit between the optimal and each other algorithms in each round on average, denoted by Δ-PoC, Δ-PoP and Δ-PoS(s).

1) Evaluation of CMAB-HS: First, when we change the total rounds N from $5{\times}10^3$ to $2{\times}10^5$ under the circumstance that numbers of sellers and selected sellers are $M{=}300$ and $K{=}10$, we evaluate the achieved total revenue and regret in Fig. 7, and Δ-PoC, Δ-PoP and Δ-PoS(s) in Fig. 8. In Fig. 7,



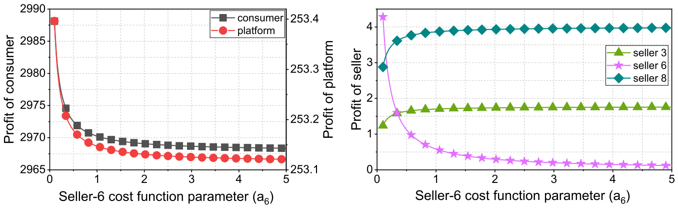
(a) Consumer vs. ω (b) All participants vs. $\omega = 1000$

Fig. 13: Profit and SE of CMAB-HS vs. p^J



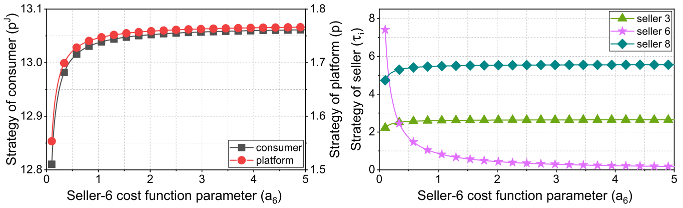
(a) Consumer and platform (b) Seller

Fig. 14: Profit vs. τ_6 and $\omega = 1000$



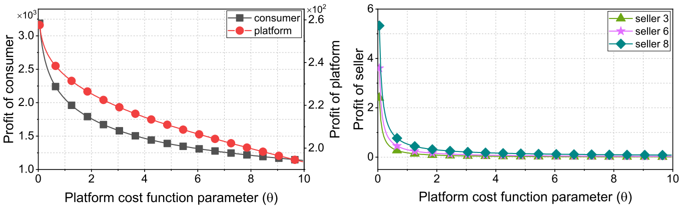
(a) Consumer and platform (b) Seller

Fig. 15: Profit vs. a_6 and $\omega = 1000$



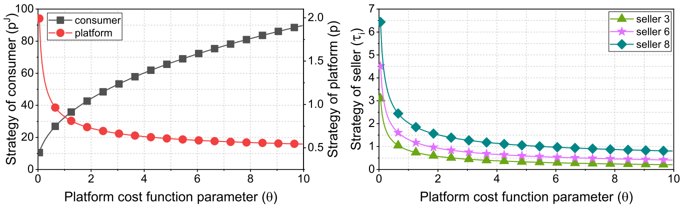
(a) Consumer and platform (b) Seller

Fig. 16: Strategy vs. a_6 and $\omega = 1000$



(a) Consumer and platform (b) Seller

Fig. 17: Profit vs. θ and $\omega = 1000$



(a) Consumer and platform (b) Seller

Fig. 18: Strategy vs. θ and $\omega = 1000$

we can observe that the total revenues of all four algorithms increase with the increasing number of total rounds N, and the two algorithms with quality learning (*i.e.*, CMAB-HS and ϵ-first) perform better than the random algorithm. The algorithms whose exploitation phases take up more rounds will achieve the near-optimal total revenues and low regrets, like in the CMAB-HS and 0.1-first algorithms. In Fig. 8, Δ-PoC, Δ-PoP and Δ-PoS(s) decrease and approach zero gradually as the number of total rounds N increases, which means that the estimation of quality is more accurate and the selection result is similar to the optimal algorithm when the number of total rounds N is large. Moreover, CMAB-HS performs better than the other two compared algorithms.

Then, we change the number of sellers M from 50 to 300 to evaluate the total revenue and regret in Fig. 9 and Δ-PoC, Δ-PoP and Δ-PoS(s) in Fig. 10 when the numbers of sellers and selected sellers are $N{=}10^5, K{=}10$. We can notice that the total revenues and regrets of all algorithms keep stable and grow very slightly as the number of sellers M increases, because the total revenues and regrets are dominated by the seller selection criteria and the selected K sellers. Since the CMAB-HS and ϵ-first algorithms always select the top-K sellers in the exploitation phase and the random algorithm always select K sellers randomly, from the ever-changing candidate sellers with various sizes, PoC, PoP and PoS(s) change slightly and keep stable in general when their profits are derived under a SE condition. Hence, Δ-PoC, Δ-PoP and Δ-PoS(s) are stable in general except for some slight fluctuations caused by the uncertainty of sensing quality when the number of sellers M increases. In addition, the two algorithms with quality learning perform better than the random algorithm. Specifically, the CMAB-HS algorithm which executes quality learning more times can obtain the near-optimal result.

Next, we let $M = 300, N = 10^5$ and change the number of selected sellers K from 10 to 60 to evaluate the total revenue, regret and average profit of each party achieved in each round (*i.e.*, average PoC, PoP and PoS(s)), as depicted in Figs. 11 and 12, respectively. In Fig. 11, the total revenues increase along with the number of selected sellers K in all algorithms, because each total revenue is accumulated based on the KN selected sellers. Moreover, the total quality estimation error becomes larger if the number of selected sellers K is large, so that the regrets also increase with the increasing K. We can notice that the regrets of algorithms with quality learning have a relatively low growth rate, where CMAB-HS performs better than other algorithms. In Figs. 12(a) and 12(b), average PoC and PoP achieved by the algorithms with quality learning in each round keep stable on the whole. However, larger K results in more cost and more low-quality sellers to be selected. Besides, more sellers selected in one round will reduce the profit of each seller, so the average PoS(s) achieved in each round decreases dramatically along with the increase of K in Fig. 12(c). Overall, the performance of CMAB-HS approaches to the optimal algorithm and is better than other compared algorithms in Fig. 12. Taking the regret and profit together into consideration, smaller K with larger N obtains a better comprehensive performance.

2) Evaluation of HS: The Hierarchical Stackelberg (HS) game is played by the consumer, the platform and the selected sellers in each round to determine their strategies. Since the decision-making process is similar in every round, we randomly select one round to evaluate the profit and strategy of individual participant. Notice that we set K=10.

First, we evaluate PoC under different consumer's data valuation parameter ω when we increase the value of SoC (*i.e.*, p^J) in Fig. 13(a). We can see that each PoC will find a maximum point (*i.e.*, SE point of HS) when p^J increases from 0 to 40, and the larger ω will harvest the larger PoC and SoC. Then, we set ω=1000 to observe the detailed change of PoC, PoP and PoS(s) of sellers $3, 6, 8$, where we use PoS-i to denote seller i's PoS. As illustrated in Fig. 13(b), PoC will find a maximum point (the SE point) but PoP and PoS(s) will continually increase as p^J increases.

Then, we evaluate the affect of changed PoS on PoC, PoP and PoS(s) if we fixed SoC and SoP as the optimal value. Both of PoC and PoP will increase at first and decrease late in Fig.14(a), which means that they can find their own maximum point in theory, respectively, even though the optimal value of SoS-6 is out of range. In our CMAB-HS mechanism, we aim to maximize all participants' profits simultaneously, *i.e.*, SE. However, the SE point may not be the absolutely maximum value for each participant if only considering itself, but it is the equilibrium to incentivize all participants to take part in the data trading and obtain satisfied profits. On the other hand, according to Eqs. (5) and (20), each PoS-i is only influenced by its quality and all sellers' cost parameters (*i.e.*, $\boldsymbol{a}, \boldsymbol{b}$) and data collection price (*i.e.*, p), so that only PoS-6 will vary with the change of SoS-6 but PoS-3 and PoS-8 will not change.

Moreover, we change the cost parameter a_6 of seller 6 to evaluate its affect on profit and strategy. It should be noticed that the larger cost of seller 6 is caused by the larger a_6, according to Eq. (6). As illustrated in Fig. 15, PoC, PoP and PoS-6 decline sharply from a_6=0 to 1, and level off gradually with the increasing a_6. Contrary to PoS-6, PoS-3 and PoS-8 will increase from a_6=0 to 1 and later flatten. Accordingly in Fig. 16(a), we can see that SoC and SoP mark a complete reversal of the profit trend, because the consumer and the platform need to raise prices (*i.e.*, p^J and p) when seller 6's cost increases. But the changing trend of SoS(s) is same with PoS(s), for the reason that the profit of each seller is positively correlated to its sensing time τ_i, as illustrated in Fig. 16(b).

Finally, we evaluate the profit and strategy when the cost parameter θ of platform varies. In Fig. 17, PoC, PoP and PoS(s) first decrease significantly and approach to a flat later, due to the incremental cost of platform with the increasing θ. It is also why the consumer needs to provide higher price p^J for the platform, and the platform will reduce price p for sellers to guarantee profit, as shown in Fig. 18(a). Since the data collection price p is lowered, each seller will reduce sensing time to maximize profit accordingly in Fig. 18(b).

VI. Related Works

Many researchers work on designing data trading mechanisms to monetize the large-scale valuable data recently, which place emphases on data management [28], data acquisition [29], [30], data trading [31], [32], etc. On one hand, data is either already existed in these works or shared by previous data owners, which cannot handle the problem of data source scarcity, *i.e.*, the data consumers cannot find any data meeting its customized demands (*e.g.*, the dynamic spatio-temporal data). On the other hand, most mechanisms do not consider the incentive issues, which will wear down the willingness of data owners to share data. Although [31], [32] pay incentive rewards to participants by contributing weight for query results, they cannot prevent the strategic participants from sacrificing system utility for their own profits. Hence, we focus on the customized data trading and incentive mechanism design, and review the related works in the following two aspects.

Incentive mechanism: Incentive mechanisms have been widely used to realize different optimization objectives. For example, [9] designs a reverse auction-based quality-aware incentive framework to incentivize seller's participation and realize social welfare maximization. [4] designs two privacy-preserving auction-based incentive mechanisms for seller selection to achieve social cost minimization. [10] proposes the reverse auction-based incentive mechanism to select reliable sellers, which can minimize the platform's total payment. However, all of these works focus on single goal optimization, which cannot be applied to our scenario. Some researches are devoted to the multiple goals optimization [16]–[21], which play two-stage non-cooperative stackelberg games between two parties to realize profit maximization of each party simultaneously, while [15] plays the cooperative game to derive Walrasian Equilibrium in data trading. Moreover, [15]–[17] use the similar quadratic function to denote cost and [16], [19]–[21] use the Piece-wise linear functions. For valuation function, [15] adopts the CobbDouglas production function and [16], [18]–[21] leverage the log functions, all of which are diminishing marginal valuation functions. Even though these works realize multiple optimizations, they do not suit for our three party game with unknown quality issues in data trading.

CMAB mechanism: Lots of researches focus on CMAB problem [25]–[27], [33], [34]. For instance, [25] extends the UCB strategy to solve the CMAB problem for online unknown worker recruitment with $O(NLK^3 \ln B)$ regret. [33] proposes a CUCB algorithm that achieves $O(\ln n)$ regret for general CMAB problems. [34] also designs a UCB-based unknown worker selection algorithm with dynamic budget allocation. However, these CMAB researches focusing on arm-pulling policy (*e.g.*, worker selection) cannot tackle the incentive issues in our data trading scenario, while only a few works consider the incentive issues [35]–[37]. In [35], the authors designs a no-regret posted price mechanism, BP-UCB, which is budget feasible and truthful. [36] proposes an auction and CMAB combined mechanism (MAB-MDR) to incentivize strategic users and obtain a sublinear $O(T^{\frac{2}{3}})$. [37] utilizes the

UCB-based approach to allocate reward and recruits workers under the reward constraint. But these CMAB mechanisms cannot be directly applied to our problem as they fail to maximize three parties' profits simultaneously. So we combine CMAB with HS to design a data trading mechanism which is effective in the scenario of three parties.

VII. Conclusion

In this paper, we focus on the problems of seller selection and incentive strategy design in practical CDT systems where the sensing qualities of sellers are unknown. We model the seller selection as a K-armed combinatorial multi-armed bandit problem and adopt hierarchical Stackelberg game to stimulate participation. We propose a data trading mechanism, called CMAB-HS, which selects sellers and determines strategies iteratively in each trading round based on the extended UCB-value. Through rigorous analysis and extensive simulations, we prove that the CMAB-HS mechanism can achieve Stackelberg Equilibrium and a tight bound on regret.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 61872330, 61572457, 61572336, 62072428, 61632016, U1709217, 61379132, 61772492), the National Key R&D Program of China (Grant No. 2018AAA0101204), the Natural Science Foundation of Jiangsu Province in China (Grant No. BK20191194, BK20131174, BK2009150), the Natural Science Research Project of Jiangsu Higher Education Institution (No. 18KJA520010), Anhui Initiative in Quantum Information Technologies (Grant No. AHY150300), and the CAS Pioneer Hundred Talents Program. Mingjun Xiao and An Liu are the corresponding authors.

References

[1] "Thingful," https://www.thingful.net/.

[2] "Thingspeak," https://thingspeak.com/.

[3] Y. Tong, Z. Zhou, Y. Zeng, L. Chen, and C. Shahabi, "Spatial crowdsourcing: a survey," *The VLDB Journal*, vol. 29, no. 1, pp. 217–250, 2020.

[4] J. Lin, D. Yang, M. Li, J. Xu, and G. Xue, "Frameworks for privacy-preserving mobile crowdsensing incentive mechanisms," *IEEE. Trans. Mob. Comput.*, vol. 17, no. 8, pp. 1851–1864, 2018.

[5] C. Jiang, L. Gao, L. Duan, and J. Huang, "Data-centric mobile crowdsensing," *IEEE. Trans. Mob. Comput.*, vol. 17, no. 6, pp. 1275–1288, 2018.

[6] W. Liu, Y. Yang, E. Wang, and J. Wu, "Dynamic user recruitment with truthful pricing for mobile crowdsensing," in *IEEE INFOCOM*, 2020.

[7] G. Yang, X. Shi, L. Feng, S. He, Z. Shi, and J. Chen, "Cedar: A cost-effective crowdsensing system for detecting and localizing drones," *IEEE. Trans. Mob. Comput.*, vol. 19, no. 9, pp. 2028–2043, 2020.

[8] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "Trading data in the crowd: Profit-driven data acquisition for mobile crowdsensing," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 486–501, 2017.

[9] H. Jin, L. Su, D. Chen, H. Guo, K. Nahrstedt, and J. Xu, "Thanos: Incentive mechanism with quality awareness for mobile crowd sensing," *IEEE. Trans. Mob. Comput.*, vol. 18, no. 8, pp. 1951–1964, 2019.

[10] H. Jin, L. Su, H. Xiao, and K. Nahrstedt, "Incentive mechanism for privacy-aware data aggregation in mobile crowd sensing systems," *IEEE-ACM Trans. Netw.*, vol. 26, no. 5, pp. 2019–2032, 2018.

[11] H. Gao, C. H. Liu, J. Tang, D. Yang, P. Hui, and W. Wang, "Online quality-aware incentive mechanism for mobile crowd sensing with extra bonus," *IEEE. Trans. Mob. Comput.*, vol. 18, no. 11, pp. 2589–2603, 2019.

[12] Z. Wang, J. Li, J. Hu, J. Ren, Z. Li, and Y. Li, "Towards privacy-preserving incentive for mobile crowdsensing under an untrusted platform," in *IEEE INFOCOM*, 2019.

[13] G. Gao, M. Xiao, J. Wu, S. Zhang, L. Huang, and G. Xiao, "Dpdt: A differentially private crowd-sensed data trading mechanism," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 751–762, 2020.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[15] X. Duan, C. Zhao, S. He, P. Cheng, and J. Zhang, "Distributed algorithms to compute walrasian equilibrium in mobile crowdsensing," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4048–4057, 2017.

[16] Y. Zhan, C. H. Liu, Y. Zhao, J. Zhang, and J. Tang, "Free market of multi-leader multi-follower mobile crowdsensing: An incentive mechanism design by deep reinforcement learning," *IEEE. Trans. Mob. Comput.*, vol. 19, no. 10, pp. 2316–2329, 2020.

[17] M. H. Cheung, F. Hou, and J. Huang, "Make a difference: Diversity-driven social mobile crowdsensing," in *IEEE INFOCOM*, 2017.

[18] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *ACM MobiCom*, 2012.

[19] B. Gu, X. Yang, Z. Lin, W. Hu, M. Alazab, and R. Kharel, "Multi-agent actor-critic network-based incentive mechanism for mobile crowdsensing in industrial systems," *IEEE Trans. Ind. Inform.*, pp. 1–1, 2020.

[20] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538–549, 2012.

[21] S. Maharjan, Y. Zhang, and S. Gjessing, "Optimal incentive design for cloud-enabled multimedia crowdsourcing," *IEEE Trans. Multim.*, vol. 18, no. 12, pp. 2470–2481, 2016.

[22] S. Yang, F. Wu, S. Tang, T. Luo, X. Gao, L. Kong, and G. Chen, "Selecting most informative contributors with unknown costs for budgeted crowdsensing," in *IEEE IWQoS*, 2016.

[23] "Chicago taxi trips," https://www.kaggle.com/chicago.

[24] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Springer ECML*, 2005.

[25] G. Gao, J. Wu, M. Xiao, and G. Chen, "Combinatorial multi-armed bandit based unknown worker recruitment in heterogeneous crowdsensing," in *IEEE INFOCOM*, 2020.

[26] H. Wang, Y. Yang, E. Wang, W. Liu, Y. Xu, and J. Wu, "Combinatorial multi-armed bandit based user recruitment in mobile crowdsensing," in *IEEE SECON*, 2020.

[27] M. T. Rashid, D. Y. Zhang, and D. Wang, "Socialdrone: An integrated social media and drone sensing system for reliable disaster response," in *IEEE INFOCOM*, 2019.

[28] J. Stoyanovich, B. Howe, and H. Jagadish, "Responsible data management," *Proc. VLDB Endow.*, vol. 13, no. 12, pp. 3474–3488, 2020.

[29] J. Bater, Y. Park, X. He, X. Wang, and J. Rogers, "Saqe: practical privacy-preserving approximate query processing for data federations," *Proc. VLDB Endow.*, vol. 13, no. 12, pp. 2691–2705, 2020.

[30] Y. Li, H. Sun, B. Dong, and H. Wang, "Cost-efficient data acquisition on online data marketplaces for correlation analysis," *Proc. VLDB Endow.*, vol. 12, no. 4, pp. 362–375, 2018.

[31] D. J. Abadi, O. Arden, F. Nawab, and M. Shadmon, "Anylog: a grand unification of the internet of things." in *CIDR*, 2020.

[32] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," *Proc. VLDB Endow.*, vol. 13, no. 12, pp. 1933–1947, 2020.

[33] Y. Xia, Q. Tao, W. Ma, N. Yu, and T. Y. Liu, "Budgeted multi-armed bandits with multiple plays," in *IJCAI*, 2016.

[34] X. Gao, S. Chen, and G. Chen, "Mab-based reinforced worker selection framework for budgeted spatial crowdsensing," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2020.

[35] A. Singla and A. Krause, "Truthful incentives in crowdsourcing tasks using regret minimization mechanisms," in *ACM WWW*, 2013.

[36] S. Jain, B. Narayanaswamy, and Y. Narahari, "A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids," in *AAAI*, 2014.

[37] H. Gao, Y. Xiao, H. Yan, Y. Tian, D. Wang, and W. Wang, "A learning-based credible participant recruitment strategy for mobile crowd sensing," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5302–5314, 2020.