INCEPTION: Incentivizing Privacy-Preserving Data Aggregation for Mobile Crowd Sensing Systems

Haiming Jin† Lu Su† Houping Xiao† Klara Nahrstedt†
†Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
‡Department of Computer Science and Engineering, State University of New York at Buffalo, NY, USA
†hjin8@illinois.edu, {lusu, houpingx}@buffalo.edu, klara@illinois.edu

ABSTRACT

The recent proliferation of human-carried mobile devices has given rise to mobile crowd sensing (MCS) systems that outsource the collection of sensory data to the public crowd equipped with various mobile devices. A fundamental issue in such systems is to effectively incentivize worker participation. However, instead of being an isolated module, the incentive mechanism usually interacts with other components which may affect its performance, such as data aggregation component that aggregates workers’ data and data perturbation component that protects workers’ privacy. Therefore, different from past literature, we capture such interactive effect, and propose INCEPTION, a novel MCS system framework that integrates an incentive, a data aggregation, and a data perturbation mechanism. Specifically, its incentive mechanism selects workers who are more likely to provide reliable data, and compensates their costs for both sensing and privacy leakage. Its data aggregation mechanism also incorporates workers’ reliability to generate highly accurate aggregated results, and its data perturbation mechanism ensures satisfactory protection for workers’ privacy and desirable accuracy for the final perturbed results. We validate the desirable properties of INCEPTION through theoretical analysis, as well as extensive simulations.

CCS Concepts

•Human-centered computing → Mobile computing; •Networks → Network economics; •Security and privacy → Privacy protections;

1. INTRODUCTION

The recent popularity of increasingly capable human-carried mobile devices (e.g., smartphones, smartglasses, smartwatches) with a plethora of on-board sensors (e.g., compass, accelerometer, gyroscope, camera, GPS) has given rise to mobile crowd sensing (MCS), a newly-emerged sensing paradigm that outsources the collection of sensory data to a crowd of participating users, namely (crowd) workers. Currently, a large variety of MCS systems [1–5] have been deployed which serve a wide spectrum of applications, including healthcare, indoor floor plan reconstruction, smart transportation, and many others.

Participating in MCS is usually costly for individual workers, since it consumes not only workers’ time but also the system resources (e.g., battery, computing power) of their mobile devices. Therefore, it is essential to design incentive mechanisms to stimulate worker participation. Typically, an incentive mechanism selects a subset of workers from the pool of potential participants to execute sensing tasks, and determines the payments to them that effectively compensate their participation costs. In real practice, an MCS system usually contains some other components which interact with the incentive mechanism and thus may affect its performance, such as data aggregation component that aggregates workers’ data and data perturbation component that protects workers’ privacy. Therefore, different from the isolated design of the incentive mechanism in past literature [6–21], we capture such interactive effect, and propose INCEPTION†, a novel MCS system framework with an integrated design of the incentive, data aggregation, and data perturbation mechanism. Below, we would like to shed some light on our design philosophy.

On one hand, the design of the incentive mechanism highly depends on how the platform aggregates workers’ data. The sensory data provided by individual workers are usually not reliable due to various factors (e.g., poor sensor quality, environmental noise, lack of sensor calibration). Therefore, the platform (i.e., a cloud-based central server) has to properly aggregate workers’ noisy and even conflicting data so as to cancel out the possible errors from individual workers. Intuitively, if workers’ data are aggregated using naive methods (e.g., average and voting) that regard all workers equally, the incentive mechanism does not need to view them differently in terms of their reliability. However, a weighted aggregation scheme that assigns higher weights to workers with higher reliability is much more favorable in that it makes

†The name INCEPTION comes from INCEtive, Privacy, and data aggregation.
the aggregated results closer to the data provided by more reliable workers. Therefore, we propose a weighted data aggregation mechanism that incorporates workers’ diverse reliabilities to calculate highly accurate aggregated results. Accordingly, we jointly design our incentive mechanism which selects workers who are more likely to provide reliable data.

On the other hand, the incentive mechanism also needs to consider the leakage of workers’ privacy, because it incurs costs which should be compensated as well. In many MCS applications, the platform usually publishes the aggregated results, which are oftentimes beneficial to the community or society, but jeopardizes workers’ privacy. Although the platform can be considered to be trusted, there exist adversaries highly motivated to infer workers’ data, which contain their sensitive and private information, from the published results. For example, publishing aggregated health data, such as treatment outcomes, improves people’s awareness about the effects of new drugs and medical devices, but poses threats to the privacy of participating patients. Geotagging campaigns provide timely and accurate localization of physical objects (e.g., automated external defibrillator, litter, pothole), however, at the risk of leaking workers’ sensitive location information. A high possibility for excessively large privacy leakage will deter workers from participating in the first place, even though they are promised to be compensated for their privacy costs. Therefore, we propose a data perturbation mechanism that reduces workers’ privacy leakage to a reasonable degree by adding carefully controlled random noises to the original aggregated results, and jointly design the incentive mechanism that compensates their costs for not only sensing but also the remaining privacy leakage.

In summary, this paper has the following contributions.

- In this paper, we propose INCEPTION, a novel MCS system framework that integrates an incentive, a data aggregation, and a data perturbation mechanism. Such an integrated design, which captures the interactive effects among these mechanisms, is much more challenging than designing them separately.
- INCEPTION has a reverse auction-based incentive mechanism that selects reliable workers and compensates their costs for both sensing and privacy leakage, which also satisfies truthfulness and individual rationality, and minimizes the platform’s total payment for worker recruiting with a guaranteed approximation ratio.
- The data aggregation mechanism of INCEPTION also incorporates workers’ reliability and generates highly accurate aggregated results.
- Its data perturbation mechanism ensures satisfactory guarantee for the protection of workers’ privacy, as well as the accuracy of the final perturbed results.

2. RELATED WORK

Game theory has been widely adopted, thus far, by the research community in the design of incentive mechanisms for MCS systems [6–21] so as to tackle workers’ strategic behaviors. Specifically, these prior work utilize either auction [12–21] or other game-theoretic models [6–11]. Although with different objectives, including maximizing social welfare [11–15] or platform’s profit [6–9, 16–19], and minimizing social cost [20] or platform’s payment [10, 21], a common property they ensure is that workers’ costs are compensated, at least in expectation. However, only workers’ sensing costs are taken into consideration by these existing work.

Different from the aforementioned prior work, we explicitly incorporate workers’ reliability and privacy costs (motivated by [22, 23]) into the incentive mechanism and provide an integrated design of the incentive, data aggregation, and data perturbation mechanism. Note that the crowd’s private information purchased by the data analyst in [22, 23] is not necessarily obtained by sensing, and thus, sensing costs are not considered by [22, 23].

One line of past literature [24–29], highly related to this paper, investigates mobile sensing systems that preserves workers’ privacy. These prior work invariably protect workers’ privacy against an untrusted platform. In contrast, the platform is trusted in our model and threats to workers’ privacy come from the adversaries outside the MCS system inferring workers’ data using the publicly available aggregated results, which cannot be tackled by the cryptography-based methods given in [24–28]. Furthermore, unlike this paper, most of these work do not consider the issue of providing incentives to workers. Another set of existing work [30–32], orthogonal to this paper, studies privacy-preserving incentive mechanisms for mobile sensing systems. These work do not consider workers’ privacy leakage caused by the public aggregated results and how it affects the design of the incentive mechanism. Instead, they protect workers’ anonymity [30, 31] or bid privacy [32] within the incentive mechanisms.

3. PRELIMINARIES

In this section, we give an overview of INCEPTION, as well as a description of the skill level model, auction model, and design objectives.

3.1 System Overview

INCEPTION is an MCS system framework consisting of a cloud-based platform and a set of N participating workers, denoted as \( N = \{ w_1, \ldots, w_N \} \). The platform hosts a set of K sensing tasks, denoted as \( T = \{ \tau_1, \ldots, \tau_K \} \), where each task \( \tau_j \in T \) requires workers to locally sense a specific object or phenomenon, and report to the platform the sensory data in the form of continuous values. Such MCS systems, collecting continuous data from the crowd, constitute a significant portion of the MCS systems currently deployed, such as geotagging campaigns that utilize workers’ GPS data to localize physical objects (e.g., automated external defibrillator, litter, pothole), and many others.

For every task \( \tau_j \in T \), the platform aggregates workers’ data into an aggregated result, denoted as \( x_{\tau_j} \), to cancel out the errors from individual workers. Every task \( \tau_j \) has a ground truth value \( x_{\tau_j} \), unknown to the platform and the workers. If worker \( w_i \) is selected to execute task \( \tau_j \), she will provide her data \( x_{\tau_j, w_i} \) to the platform. We assume that \( x_{\tau_j} \) and \( x_{\tau_j, w_i} \)'s are normalized values within the range \([0, 1]\) for simplification of presentation. We define matrix \( x = [x_{\tau_j, w_i}] \in (\{0, 1\} \cup \{\perp\})^{N \times K} \) containing all workers’ data, where \( x_{\tau_j, w_i} = \perp \) means that task \( \tau_j \) is not executed by worker \( w_i \).

In our model, the platform publishes the aggregated results (e.g., locations of automated external defibrillators, litter, potholes) to the community or society. However, directly publishing them impairs workers’ privacy. Therefore, the platform publishes the perturbed results after adding random noises to the original ones, and ensures \( \epsilon \)-differential privacy defined in Definition 1 (adapted from [33]).
Definition 1 (Differential Privacy). We denote \( M : \{(0, 1) \cup \{ \perp \}\}^{N \times K} \rightarrow R^{K \times 1} \) as a mechanism that maps any input data matrix to a perturbed result vector. Then, the mechanism \( M \) is \( \epsilon \)-differentially private if and only if for any two data matrices \( \mathbf{x} \) and \( \mathbf{x}' \) that differ in only one entry and any \( A \subseteq R^{K \times 1} \), we have
\[
Pr[M(\mathbf{x}) \in A] \leq \exp(\epsilon)Pr[M(\mathbf{x}') \in A],
\]
where \( \epsilon \) is a small positive number usually referred to as privacy budget.

The framework of INCEPTION is illustrated in Figure 1, and its workflow is described as follows.

![Figure 1: Framework of INCEPTION](image)

- Firstly, the platform announces the set of sensing tasks \( T \) and an upper bound of the privacy budget \( \epsilon \), then picks a worker \( w_i \) (step (1)).
- Incentive Mechanism. Then, the platform starts the reverse auction-based incentive mechanism, where it acts as the auctioneer, to purchase data from participating workers, who act as bidders. Every worker \( w_i \) submits his bid \( b_i = (\Gamma_i, b_i^\epsilon, b_i^0) \) which is a triple containing the set of sensing tasks \( \Gamma_i \), which she wants to execute, as well as her bidding prices for executing them \( b_i^\epsilon \) and unit privacy loss \( b_i^0 \) (step (2)). Based on workers’ bids, the platform determines the set of winners \( S \subseteq N \) and the payment \( p_i \) to each winner \( w_i \) (step (3)). Lossers of the auction do not execute tasks and receive no payments. We denote workers’ bids and payment profile as \( \mathbf{b} = (b_1, \ldots, b_N) \) and \( \mathbf{p} = (p_1, \ldots, p_N) \), respectively.
- Data Aggregation Mechanism. Next, the platform collects winners’ sensory data (step (4)) and calculates an aggregated result \( x_i \) for each task \( \tau_i \) (step (5)).
- After collecting workers’ data, the platform pays workers according to \( \mathbf{p} \) and reveals to them the exact value of the privacy budget \( \epsilon \) (step (6)), such as \( \epsilon = 0.25 \). The design rationale for keeping the exact value of \( \epsilon \) confidential to workers at the bidding stage and revealing it together with the payments is described in detail in Section 4.2.3.
- Data Perturbation Mechanism. Finally, the platform adds random noises to the original aggregated results and publishes the perturbed ones (step (7)). We use \( \tilde{x}_i \) to denote the perturbed result for task \( \tau_i \).

3.2 Skill Level Model

Before task \( \tau_i \) is executed by worker \( w_i \), her data about this task can be regarded as a random variable \( X_{i,j} \). Then, we define a worker’s skill level in Definition 2.

Definition 2. Worker \( w_i \)’s skill level \( \theta_{i,j} \) for task \( \tau_j \) is defined as the expected absolute difference between her data and the ground truth, i.e.,
\[
\theta_{i,j} = E[|X_{i,j} - x_j^*|] \in [0, 1],
\]
where the expectation is taken over the randomness of \( X_{i,j} \).

We use \( \Theta = [\theta_{i,j}] \in [0, 1]^{N \times K} \) to denote the skill level matrix of all workers.

We assume that the skill level matrix \( \Theta \) is a priori known to the platform. In practice, the platform can keep a historical record of \( \Theta \), which can be obtained by many methods. For example, since a worker’s skill levels for similar tasks typically tend to be similar, the platform could assign some tasks with known ground truths to workers and utilize workers’ sensory data about these tasks to estimate their skill levels for similar tasks as in [34]. In scenarios where ground truths are not available, \( \Theta \) can still be effectively estimated utilizing workers’ previously submitted sensory data about similar tasks by algorithms proposed in [35, 36] or inferred from some of workers’ characteristics (e.g., a worker’s reputation and experience for similar tasks, the price of a worker’s sensors) using the methods in [37].

3.3 Auction Model

In this paper, as in most prior work, we assume that workers are selfish and strategic that aim to maximize their own utilities. We use the term bundle to refer to any subset of the overall task set \( T \) in the rest of this paper. Since every worker bids on one bundle of tasks in the INCEPTION framework, we model the incentive mechanism as a single-minded reverse combinatorial auction. However, different from the traditional combinatorial auction [38], we study the scenario where workers explicitly consider privacy leakage as one of the sources for their costs. Therefore, we propose the single-minded reverse combinatorial auction with privacy cost (pSRC auction), formally defined in Definition 3, as the incentive mechanism.

Definition 3 (pSRC Auction). In a single-minded reverse combinatorial auction with privacy cost (pSRC auction), each worker \( w_i \) has only one interested bundle \( \Gamma_i^* \). Her cost of executing the bundle of tasks, namely sensing cost, is denoted as \( C_i^s(\epsilon) \) (unknown to the platform). Additionally, she has a cost for privacy leakage, namely privacy cost, denoted as \( C_i^\epsilon(\Gamma_i^*, \epsilon) \), if \( \epsilon \)-differential privacy is guaranteed. Hence, worker \( w_i \)’s cost function is defined as in Equation (3).
\[
C_i(\Gamma, \epsilon) = \begin{cases} 
C_i^s(\epsilon) + C_i^\epsilon(\Gamma_i^*, \epsilon), & \text{if } \Gamma \subseteq \Gamma_i^* \\
\infty, & \text{otherwise}.
\end{cases}
\]

For the tasks that do not belong to worker \( w_i \)’s interested bundle \( \Gamma_i^* \), either she is not able to execute them or executing these tasks incurs a large cost. Therefore, we assign a \( +\infty \) cost to these tasks in Equation (3).

A major difference between the cost function defined in Equation (3) and those in prior work [7–21] is that the privacy cost \( C_i^\epsilon(\epsilon) \) is explicitly integrated into it. Such integration is reasonable and necessary. In an MCS system where the platform utilizes a worker’s private and sensitive data in a way that incurs privacy leakage, the worker will not be effectively incentivized to participate unless both her sensing and privacy cost are compensated. For any worker \( w_i \), the privacy cost \( C_i^\epsilon(\epsilon) \) is positively correlated with the privacy budget \( \epsilon \), because \( \epsilon \) in fact captures the amount of privacy leakage of the MCS system. Therefore, we adopt the natural linear model for privacy cost as in [22, 23] where \( C_i^\epsilon(\epsilon) = c^\epsilon_i \epsilon \) with \( c^\epsilon_i \) representing worker \( w_i \)’s cost for unit privacy leakage. Similar to \( c^s_i \), \( c^\epsilon_i \) is also unknown to the platform. Next, we define a worker’s utility in Definition 4.

Definition 4 (Worker’s Utility). Any worker \( w_i \)’s utility \( u_i \) is defined as
\[
u_i(\mathbf{p}, \mathbf{x}) = \sum_{j=1}^{K} u_i^j(\mathbf{p}, \mathbf{x}),
\]
where \( u_i^j(\mathbf{p}, \mathbf{x}) \) is defined as in Equation (4) for each task \( \tau_j \) in bundle \( \Gamma_i^* \). We consider the platform as a social planner who pays workers for the bundle of tasks they execute and takes into account the workers’ privacy leakages as well as the sensing costs.
is defined as
\[ u_i = \begin{cases} p_i - c_i^l - c_i^r \epsilon, & \text{if } u_i \in S \\ 0, & \text{otherwise} \end{cases} \quad (4) \]

Apart from workers’ utilities, we are also interested in the platform’s total payment defined in Definition 5.

**Definition 5 (Platform’s Total Payment).** Given the payment profile \( p \) and the winner set \( S \), the platform’s total payment is \( P = \sum_{i:w_i \in S} p_i \).

### 3.4 Design Objective

In this section, we aim to ensure that INCEPTION bears the following desirable properties.

Since workers are strategic in our model, it is possible that any worker \( w_i \) submits a bid \((\Gamma_i, b_i^l, b_i^r)\) that deviates from the true value \((\Gamma_i^*, c_i^l, c_i^r)\). However, one of our objectives is to design a truthful incentive mechanism defined in Definition 6.

**Definition 6 (Truthfulness).** A pSRC auction is truthful if and only if bidding the true value \((\Gamma_i^*, c_i^l, c_i^r)\) is the dominant strategy for each worker \( w_i \), i.e., bidding \((\Gamma_i^*, c_i^l, c_i^r)\) maximizing each worker \( w_i \)’s utility for all possible values of other workers’ bids and the privacy budget \( \epsilon \).

By Definition 6, we aim to ensure the truthful bidding of the interested bundle \( \Gamma_i^* \), the sensing cost \( c_i^l \), and the cost for unit privacy leakage \( c_i^r \) for every worker \( w_i \). Apart from truthfulness, another desirable and necessary property is individual rationality defined in Definition 7.

**Definition 7 (Individual Rationality).** A pSRC auction is individual rational if and only if no worker receives negative utility, i.e., we have \( u_i \ge 0 \) for each worker \( w_i \).

Individual rationality in our pSRC auction means that a worker’s sensing and privacy cost are both compensated, which is crucial to effectively incentivize worker participation. As mentioned in Section 3.1, we aim to design an MCS system that ensures \( \epsilon \)-differential privacy. However, the perturbation added to the aggregated results impairs their accuracy which is mathematically defined in Definition 8.

**Definition 8 ((\( \alpha, \beta \))-Accuracy).** For two random variables \( Y_1 \) and \( Y_2 \) within the range \([0, 1] \), \( Y_1 \) is \((\alpha, \beta)\)-accurate to \( Y_2 \) if and only if \( \Pr[Y_1 - Y_2 \ge \alpha] \le \beta \), where \( \alpha, \beta \in (0, 1) \). Note that \( Y_2 \) could also be a constant.

We use \( X_j \) to denote the random variable corresponding to \( x_j \) (i.e., the perturbed result for task \( \tau_j \)). Facing the tradeoff between privacy and accuracy, we also need to carefully control the amount of noises added to the aggregated results and ensure that \( X_j \) is \((\alpha, \beta)\)-accurate to the ground truth \( x_j^* \) for every task \( \tau_j \) with sufficiently small \( \alpha \) and \( \beta \) within \((0, 1)\). That is, we aim to ensure that the perturbed results are fairly close to ground truths with high probability.

In short, our objective is to design a differentially private MCS system that provides satisfactory accuracy guarantee for the final perturbed results, and incentivizes worker participation in a truthful and individual rational manner.

### 4. DESIGN DETAILS

In this section, we provide our design details for the incentive, data aggregation, and data perturbation mechanism.

#### 4.1 Data Aggregation Mechanism

##### 4.1.1 Proposed Mechanism

Although the data aggregation mechanism comes after the incentive mechanism in INCEPTION’s workflow, we introduce it first, as it affects the design of incentive mechanism.

To guarantee that the perturbed results have satisfactory accuracy, the original aggregated results before perturbation need to be accurate enough in the first place. Therefore, we reasonably assume that the platform uses a weighted aggregation method to calculate the aggregated result \( x_j \) for each task \( \tau_j \) based on workers’ data. That is, given the winner set \( S \) determined by the incentive mechanism, we have

\[ x_j = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} x_{i,j}, \quad (5) \]

where \( \lambda_{i,j} > 0 \) is the weight of worker \( w_i \) on task \( \tau_j \) with \( \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} = 1 \) for every task \( \tau_j \).

The motivation for utilizing weighted aggregation is to capture the effect of workers’ diverse skill levels on the calculation of the aggregated results. Intuitively, we should assign higher weights to workers whose sensory data are more likely to be close to the ground truths, which makes the aggregated results closer to the data provided by more reliable workers. In fact, many state-of-the-art data aggregation methods [35, 36] utilize such weighted aggregation to calculate the aggregated results. Since the accuracy of the aggregated results highly depends on how exactly the weight \( \lambda_{i,j} \)’s are chosen, we propose the following data aggregation mechanism in Algorithm 1.

**Algorithm 1: Data Aggregation Mechanism**

**Input:** \( \alpha, \theta, b, x, S \);

**Output:** \((x_1, \ldots, x_K)\);

1. **foreach** \( j \) s.t. \( \tau_j \in \mathcal{T} \) **do**

   \[ x_j \leftarrow \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \frac{(\alpha_j - \theta_{i,j}) x_{i,j}}{\sum_{k:w_k \in S, \tau_j \in \Gamma_k} (\alpha_j - \theta_{k,j})}; \]

2. **return** \((x_1, \ldots, x_K)\);

Besides the skill level matrix \( \theta \), the bid profile \( b \), workers’ data \( x \), and the winner set \( S \), Algorithm 1 also takes as input a vector of positive real numbers \( \alpha = (\alpha_1, \ldots, \alpha_K) \) chosen by the platform with \( \max_{i,j} \in \Gamma_i \theta_{i,j} < \alpha_j < 0.5 \). Note that large \( \theta_{i,j} \) indicates low reliability, and any worker \( w_i \) with \( \theta_{i,j} < 0.5 \) will not be selected by the incentive mechanism to execute task \( \tau_j \). The aggregated result \( x_j \) for every task \( \tau_j \) is calculated (line 2) using Equation (5) with the weight

\[ \lambda_{i,j} = \frac{\sum_{k:w_k \in S, \tau_j \in \Gamma_k} (\alpha_j - \theta_{k,j})}{\sum_{k:w_k \in S, \tau_j \in \Gamma_k} (\alpha_j - \theta_{k,j})}, \quad \forall w_i \in S, \tau_j \in \Gamma_i. \quad (6) \]

By Equation (6), worker \( w_i \)’s weight for task \( \tau_j \), namely \( \lambda_{i,j} \), increases with the decrease of \( \theta_{i,j} \). Such a design choice conforms to our intuition that the less the expected deviation of worker \( w_i \)’s data compared to the ground truth \( x_j^* \), the more \( x_{i,j} \) should be counted in the calculation of the aggregated result \( x_j \). Formal analysis about the data aggregation mechanism is provided in Section 4.1.2.

#### 4.1.2 Analysis

In Theorem 1, we prove that the aggregated result \( x_j \) calculated using Algorithm 1 guarantees desirable accuracy compared to the ground truth \( x_j^* \).

**Theorem 1.** We use \( X_j \) to denote the random variable representing the aggregated result for task \( \tau_j \). The data aggregation mechanism proposed in Algorithm 1 minimizes the upper bound of the probability \( \Pr[[X_j - x_j^*] \ge \alpha_j] \) and ensures that for every task \( \tau_j \), we have

\[ \Pr[[X_j - x_j^*] \ge \alpha_j] \le \exp \left( -2 \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \right). \quad (7) \]
Proof. From Equation (5), we have
\[ |X_j - x_j^*| = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} x_{i,j} - x_j^* \]
\[ = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (x_{i,j} - x_j^*) \]
\[ \leq \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (x_{i,j} - x_j^*) \]
\[ \leq \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (x_{i,j} - x_j^*) \]

We define a random variable \( Y_j \) for every task \( \tau_j \) as \( Y_j = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (x_{i,j} - x_j^*) \), which is the sum of random variables \( Y_{i,j} \)'s with \( Y_{i,j} = [\lambda_{i,j} (x_{i,j} - x_j^*)] \in [0, \lambda_{i,j}] \). Thus,
\[ E(Y_j) = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} E[(x_{i,j} - x_j^*)] = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} \theta_{i,j} \]

Therefore, from the Chernoff-Hoeffding bound, we have
\[ \Pr[|X_j - x_j^*| \geq \alpha_j] \leq \Pr \left[ \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (x_{i,j} - x_j^*) \geq \alpha_j \right] \]
\[ = \Pr[\lambda \geq \alpha_j] \]
\[ = \Pr[\lambda - E(\lambda) > \alpha_j - E(\lambda)] \]
\[ \leq \exp \left( -\frac{2(\alpha_j - E(\lambda))^2}{\sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2} \right) \]
\[ = \exp \left( -\frac{2(\alpha_j^2 - \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2 \theta_{i,j})}{\sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2} \right) \]
\[ = \exp \left( -\frac{2(\sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2 \alpha_j - \theta_{i,j})^2}{\sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2} \right) \]
\[ \leq \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \]

We denote the vector \( \lambda_j \) as \( \lambda_j = [\lambda_{i,j}] \) for every task \( \tau_j \) containing every \( \lambda_{i,j} \) such that \( w_i \in S \) and \( \tau_j \in \Gamma_i \). Therefore, minimizing the upper bound of \( \Pr[|X_j - x_j^*| \geq \alpha_j] \) is equivalent to maximizing the function \( \varphi(\lambda_j) \) defined as
\[ \varphi(\lambda_j) = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} (\alpha_j - \theta_{i,j})^2 \]

From the Cauchy-Schwarz inequality, we have that
\[ \varphi(\lambda_j) \leq \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j}^2 \left( \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \right) \]
\[ = \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \]

and equality is achieved when \( \lambda_{i,j} \propto \alpha_j - \theta_{i,j} \).

Using the fact that \( \sum_{i:w_i \in S, \tau_j \in \Gamma_i} \lambda_{i,j} = 1 \), we have
\[ \lambda_{i,j} = \frac{\alpha_j - \theta_{i,j}}{\sum_{k:w_k \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{k,j})} \] (8)

Therefore, when \( \lambda_{i,j} \)'s satisfy Equation (8), we have
\[ \Pr[|X_j - x_j^*| \geq \alpha_j] \leq \exp \left( -2 \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \right) \]

which is exactly the Equation (7) in Theorem 1.

By Theorem 1, the data aggregation mechanism proposed in Algorithm 1 upper bounds the probability of \( \Pr[|X_j - x_j^*| \geq \alpha_j] \) by \( \exp \left( -2 \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \right) \) which is in fact the minimum upper bound for this probability. Then, we introduce Corollary 1 which is directly utilized in the design of the incentive mechanism in Section 4.2.

**Corollary 1.** For the data aggregation mechanism proposed in Algorithm 1, if
\[ \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \geq \frac{1}{2} \ln \left( \frac{1}{\beta_j} \right) \] (9)
then we have \( \Pr[|X_j - x_j^*| \geq \alpha_j] \leq \beta_j \), where \( \beta_j \in (0, 1) \) for every task \( \tau_j \), is a parameter chosen by the platform. We use \( \beta \) to denote the vector \( \beta_j \).

**Proof.** Corollary 1 directly follows from Theorem 1. If we let the upper bound of \( \Pr[|X_j - x_j^*| \geq \alpha_j] \) guaranteed by Algorithm 1 to be no greater than \( \beta_j \in (0, 1) \), we have
\[ \exp \left( -2 \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \right) \leq \beta_j, \]
which is equivalent to exactly
\[ \sum_{i:w_i \in S, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 \geq \frac{1}{2} \ln \left( \frac{1}{\beta_j} \right). \] (10)

Therefore, together with Theorem 1, we know that Inequality (10) implies \( \Pr[|X_j - x_j^*| \geq \alpha_j] \leq \beta_j. \)

**Corollary 1** states that \( \alpha_j, \beta_j \)-accurate is guaranteed for the aggregated result of task \( \tau_j \) compared to the ground truth \( x_j^* \), if the condition specified by Inequality (9) is satisfied for the set of selected winners \( S \) in the incentive mechanism proposed in Section 4.2.

### 4.2 Incentive Mechanism

In this section, we introduce the mathematical formulation, design details and the analysis of the proposed incentive mechanism.

#### 4.2.1 Mathematical Formulation

As mentioned in Section 3.3, our incentive mechanism is based on the pSRC auction defined in Definition 3. In this paper, we aim to design a pSRC auction that minimizes the platform’s total payment with satisfactory data aggregation accuracy. Such a design choice exactly captures the objective of most MCS systems, that is to collect high quality data from the crowd with minimum total expense. The formal mathematical formulation is given in the following pSRC auction total payment minimization (pSRC-TPM) problem.

**pSRC-TPM Problem:**
\[
\min \sum_{i \in N} p_i y_i \tag{11}
\]
\[
\text{s.t.} \quad \sum_{i \in N, \tau_j \in \Gamma_i} (\alpha_j - \theta_{i,j})^2 y_i \geq \frac{1}{2} \ln \left( \frac{1}{\beta_j} \right), \quad \forall \tau_j \in \mathcal{T} \tag{12}
\]
\[
y_i \in \{0, 1\}, \quad p_i \in [0, +\infty), \quad \forall w_i \in N \tag{13}
\]

**Constants.** The pSRC-TPM problem takes as inputs the worker set \( \mathcal{N} \), the task set \( \mathcal{T} \), workers’ bid profile \( \mathbf{b} \), the skill level matrix \( \Theta \), and the \( \beta \) and \( \alpha \) vector.

**Variables.** The pSRC-TPM problem has a vector of \( N \) binary variables, denoted as \( \mathbf{y} = (y_1, \ldots, y_N) \). The variable \( y_i = 1 \) indicates that the worker \( w_i \) is selected as a winner (i.e., \( w_i \in S \)); otherwise \( w_i \notin S \). The second vector of variables is the payment profile \( \mathbf{p} = (p_1, \ldots, p_N) \), where every element takes a non-negative real value.

**Objective function.** The objective function given by
\[ \sum_{w_i \in N} p_i y_i = \sum_{i \in N} p_i \] is exactly the total payment made by the platform to all winners.

**Constraints.** Constraint (12) is equivalent to Inequality (9) given in Corollary 1, which specifies the condition that the selected winners should satisfy. By Corollary 1, any solution \( \mathbf{y} \) to the pSRC-TPM problem gives a winner set \( S \) which ensures that the aggregated result of every task \( \tau_j \) is \( (\alpha_j, \beta_j) \)-accurate to the ground truth \( x_j^* \).

To simplify presentation, we introduce the following extra notations, \( q_{i,j} = (\alpha_j - \theta_{i,j})^2 \), \( q = [q_{i,j}] \in [0, +\infty)^{N \times K} \), \( Q_j = \frac{1}{2} \ln \left( \frac{1}{\beta_j} \right) \), and \( Q = [Q_j] \in [0, +\infty)^{K \times 1} \). Therefore, Constraint (12) is simplified as
\[ \sum_{i:w_i \in S, \tau_j \in \Gamma_i} q_{i,j} y_i \geq Q_j, \quad \forall \tau_j \in \mathcal{T}. \] (14)

Besides Constraint (12), any solution to the pSRC-TPM problem should also satisfy two other inherent constraints,
namely truthfulness and individual rationality, which are difficult to formulate mathematically.

In Theorem 2, we prove the NP-hardness of the pSRC-TPM problem.

**Theorem 2.** The pSRC-TPM problem is NP-hard.

**Proof.** We consider a special case of the pSRC-TPM problem with a constant payment profile $\mathbf{p}$. With constant $p_i$’s, it becomes a binary linear program (BLP). We prove the NP-hardness of the BLP by a polynomial-time reduction from the minimum weight set cover (MWSC) problem.

The reduction starts from an instance of the NP-complete MWSC problem with a universe $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$ and a set of subsets of $\mathcal{T}$ defined as $\mathcal{R} = \{\Gamma_1, \ldots, \Gamma_N\}$. Each set $\Gamma_i \in \mathcal{R}$ has a non-negative weight $w_i$. The objective of the MWSC problem is to find the subset of $\mathcal{R}$ with the minimum total weight whose union equals to $\mathcal{T}$. We transform $\Gamma_i$ to $\Gamma'_i$, where each element $\tau_j \in \Gamma_i$ has $a_{i,j} \in \mathbb{Z}^+$ copies and require each $\tau_j$ to be covered for exactly $A_j \in \mathbb{Z}$ times. By now, an instance of the BLP with $\mathbf{q} = [a_{i,j}] \in (\mathbb{Z}^+)^{N \times K}$, $\mathbf{Q} = [A_j] \in (\mathbb{Z}^+)^{K \times 1}$, and payment profile $\mathbf{p}$ has been constructed. Actually, a richer family of problems can be represented by the BLP because elements in $\mathbf{q}$ and $\mathbf{Q}$ can be any positive real numbers besides positive integers. Hence, every instance of the MWSC problem is polynomial-time reducible to the BLP, which proves its NP-hardness. Furthermore, because the BLP is only a special case of the pSRC-TPM problem, the pSRC-TPM problem is also NP-hard.

### 4.2.2 Proposed Mechanism

Because of the NP-hardness of the pSRC-TPM problem proved in Theorem 2, directly solving it to obtain the winner set $\mathcal{S}$ and the payment profile $\mathbf{p}$ is computationally intractable when the cardinality of $\mathcal{N}$ and $\mathcal{T}$ become large. Therefore, we propose our own winner determination and pricing algorithm for the pSRC auction in Algorithm 2 and 3, respectively. The proposed algorithms are computationally efficient and approximately minimize the platform’s total payment with a guaranteed approximation ratio.

The inputs of the winner determination algorithm given in Algorithm 2 include the privacy budget $\epsilon$, bid profile $\mathbf{b}$, bid profile $\mathbf{q}$ matrix, $\mathbf{Q}$ vector, worker set $\mathcal{N}$, and task set $\mathcal{T}$. Firstly, it initializes the worker set $\mathcal{S}$ as $\emptyset$ and the residual vector of $\mathbf{Q}$, namely $\mathbf{Q}^*$ as $\mathbf{Q}$ (line 1). Then, the main loop (line 2-7) calculates the winner set $\mathcal{S}$. It is executed until the winner set $\mathcal{S}$ makes the pSRC-TPM problem feasible (line 2). We define worker $w_i$’s virtual bidding price as $b_i^v = b_i^* + b_i^p \epsilon$. In each iteration, Algorithm 2 finds the worker $w_i$ with the minimum bidding price effectiveness (line 3) defined as the ratio between her virtual bidding price and her contribution to the improvement of the feasibility of Constraint (12). Next, $w_i$ is included into the winner set $\mathcal{S}$ (line 4) and excluded from the worker set $\mathcal{N}$ (line 5). Finally, the $\mathbf{Q}^*$ vector is updated (line 6-7) before the start of the next iteration.

Apart from the same inputs taken by Algorithm 2, the pricing algorithm given in Algorithm 3 also uses the winner set $\mathcal{S}$ calculated by Algorithm 2. Firstly, it initializes the payment profile $\mathbf{p}$ as a vector of $N$ zeros (line 1). Then, the main loop (line 2-7) calculates the payment to each worker. For each worker $w_i \in \mathcal{S}$, Algorithm 2 is executed on the worker set containing all workers except $w_i$ until the point after which $w_i$ will never be selected as a winner (line 3). The winner set at this point is recorded as $\mathcal{S}'$ (line 4). For each worker $w_k \in \mathcal{S}'$, Algorithm 3 calculates worker $w_i$’s maximum virtual bidding price $b_{i,k}^v$ that makes her substitute $w_k$ as the winner. To achieve this, $b_{i,k}^v$ should satisfy

$$\sum_{j: \tau_j \in \Gamma_i} \min(Q_j^*, q_{i,j}) = \sum_{j: \tau_j \in \Gamma_k} \min(Q_j^*, q_{k,j}),$$

which is equivalent to

$$b_{i,k}^v = \frac{b_i^* + b_i^p \epsilon}{\sum_{j: \tau_j \in \Gamma_i} \min(Q_j^*, q_{i,j})} \cdot \frac{\sum_{j: \tau_j \in \Gamma_k} \min(Q_j^*, q_{k,j})}{b_i^* + b_i^p \epsilon}.$$

Then, the maximum value among these $b_{i,k}^v$’s is chosen as the payment $p_i$ to worker $w_i$ (line 7).

### 4.2.3 Analysis

Firstly, we analyze the truthfulness of the proposed pSRC auction in Theorem 3.

**Theorem 3.** The proposed pSRC auction is truthful.

**Proof.** Firstly, we fix the privacy budget $\epsilon$ and assume a worker $w_i$ wins the auction by bidding $b_i = (\Gamma_i, \mathbf{b}_i, \mathbf{b}_i^p)$. We show that the pSRC auction satisfies the properties of monotonicity and critical payment in terms of the bidding bundle $\Gamma_i$ and virtual bidding price $b_i^v = b_i^* + b_i^p \epsilon$.

- **Monotonicity.** Consider worker $w_i$’s bid $\tilde{b}_i = (\Gamma_i, \tilde{\mathbf{b}}_i, \mathbf{b}_i^p)$ with $\Gamma_i \supset \Gamma_i$ and $\tilde{b}_i^v = b_i^* + \tilde{b}_i^p \epsilon < b_i^v$. Algorithm 2 selects winners in an increasing order of the bidding price effectiveness. Hence, $\tilde{b}_i$ will also make worker $w_i$ a winner, as it increases her priority of winning compared to $b_i$.

- **Critical payment.** Algorithm 3 in fact pays every winner the supremum of all virtual bidding prices that can still make her a winner, namely critical payment. As proved in [13, 38], the monotonicity and critical pay-
straints, and variables
Theorem 5. The pSRC auction is truthful.

Proof. By Definition 4, losers of the auction receive zero utilities. From Theorem 3, every winner \( w_i \) bids to the platform the true value \( (c_i^1, c_i^2) \) and the payment \( p_i \) to this winner is exactly the supremum of all virtual bidding prices for her to win the auction. Therefore, it is guaranteed that \( p_i \geq c_i^1 + c_i^2 \epsilon \), which is equivalent to \( u_i \geq 0 \). Hence, the proposed pSRC auction is individual rational.

In our INCEPTION framework, the platform reveals the exact value of the privacy budget \( \epsilon \) when workers receive their payments so that they can evaluate their utilities after participating and confirm that their utilities are in fact non-negative. Next, we analyze the algorithmic properties of the pSRC auction.

Theorem 5. The computational complexity of the proposed pSRC auction is \( O(N^3 + N^2K) \).

Proof. The main loop (line 2-7) of Algorithm 2 terminates in worst case after \( N \) iterations. In every iteration, it takes \( O(N) \) time to find the worker with the minimum bidding price effectiveness (line 3), and at most \( K \) other iterations are needed to update the \( Q' \) vector (line 6-7). Therefore, the computational complexity of Algorithm 2 is \( O(N^3 + NK) \).

Furthermore, the computational complexity of Algorithm 3 is \( O(N^3 + N^2K) \), because there is one more layer of loop that executes for \( N \) iterations in worst case. In conclusion, the computational complexity of the pSRC auction is \( O(N^3 + N^2K) \).

Before analyzing the approximation ratio of the platform’s total payment generated by the pSRC auction to the optimal total payment, we introduce Lemma 1 and 2 that are utilized in the analysis. The two lemmas are directly related to the pSRC auction social cost minimization (pSRC-SCM) problem defined as follows.

pSRC-SCM Problem:

\[
\begin{align*}
\text{min } & \sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) y_i \\
\text{s.t. } & \sum_{i:w_i \in S, \tau_j \in \Gamma_i} q_{i,j} y_i \geq Q_j, \quad \forall \tau_j \in \mathcal{T} \\
& y_i \in \{0, 1\}, \quad \forall w_i \in \mathcal{N}
\end{align*}
\]

(15)

(16)

(17)

The pSRC-SCM problem has the same set of inputs, constraints, and variables \( \mathbf{y} = \{y_1, \cdots, y_N\} \) as the pSRC-TPM problem. Instead of the platform’s total payment, it minimizes the social cost, i.e., \( \sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) \), which is the sum of all winners’ costs.

Lemma 1. For any individual rational pSRC auction, the optimal social cost of the pSRC-SCM problem, denoted as \( \text{C}_{OPT} \), is a lower bound of the optimal total payment of the pSRC-TPM problem, denoted as \( P_{OPT} \).

Proof. Suppose \( (\mathbf{y}^*, \mathbf{p}^*) \) is the optimal solution to the pSRC-TPM problem. We have \( P_{OPT} = \sum_{i:w_i \in S} p_i y_i^* \).

Since the two problems have the same set of constraints, \( \mathbf{y}^* \) is also feasible to the pSRC-SCM problem. Furthermore, from individual rationality, we have \( p_i^* \geq (c_i^1 + c_i^2 \epsilon) y_i^* \) for every worker \( w_i \). Therefore, we have

\[
\text{C}_{OPT} \leq \sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) y_i^* \leq \sum_{i:w_i \in S} p_i^* y_i^* = P_{OPT}
\]

That is, \( \text{C}_{OPT} \) is a lower bound of \( P_{OPT} \) for any individual rational pSRC auction.

Then, we introduce Lemma 2 which is borrowed from [13] (Theorem 5 in [13]) with some minor adaptations. Similar to [13], we introduce the following notations including \( \gamma = \max_{i,j:w_i \in \mathcal{N}, \tau_j \in \mathcal{T}} (c_i^1 + c_j^2 \epsilon) q_{i,j} / |\Gamma_i| \) and \( m = \frac{1}{\Delta q} \sum_{i: \tau_j \in \mathcal{T}} Q_j \), where \( \Delta q \) is the unit measure of elements in \( \mathbf{q} \) and \( \mathbf{Q} \).

Lemma 2. The social cost generated by Algorithm 2 satisfies a \( 2\gamma H_m \)-approximation to the optimal social cost, i.e.,

\[
\sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) \leq 2\gamma H_m \text{C}_{OPT}
\]

where \( H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m} \).

The proof to Lemma 2, which can be found in [13], is omitted in this paper. We define \( \delta = \max_{i,k,w_i \in \mathcal{N}} (c_i^1 + c_k^2 \epsilon) \) and introduce the following Theorem 6 regarding the approximation ratio of the proposed pSRC auction in terms of the platform’s total payment.

Theorem 6. The platform’s total payment generated by the proposed pSRC auction satisfies a \( 2\delta \gamma H_m \)-approximation to the optimal total payment, i.e.,

\[
\sum_{i:w_i \in S} p_i \leq 2\delta \gamma H_m P_{OPT}
\]

Proof. Based on Algorithm 3, for every winner \( w_k \) there exists some worker \( w_{k_i} \) such that

\[
p_i = (c_i^1 + c_i^2 \epsilon) \frac{\sum_{j: \tau_j \in \Gamma_{k_i}} \min \{Q_j, q_{i,j}\}}{\sum_{j: \tau_j \in \Gamma_{k_i}} \min \{Q_j, q_{i,j}\}}
\]

where \( Q_j \) denotes the element corresponding to task \( \tau_j \) in the \( Q' \) vector determined on line 6 of Algorithm 3 when the worker \( w_{k_i} \) is selected as a winner. Therefore, we have

\[
\sum_{i:w_i \in S} p_i = \sum_{i:w_i \in S} \left( c_i^1 + c_i^2 \epsilon \right) \frac{\sum_{j: \tau_j \in \Gamma_{k_i}} \min \{Q_j, q_{i,j}\}}{\sum_{j: \tau_j \in \Gamma_{k_i}} \min \{Q_j, q_{i,j}\}} \leq m |\mathcal{S}| \max_{i:k \in \mathcal{N}} (c_i^1 + c_k^2 \epsilon)
\]

(18)

Furthermore, the social cost satisfies that

\[
\sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) \geq |\mathcal{S}| \min_{i,w_i \in \mathcal{N}} (c_i^1 + c_i^2 \epsilon)
\]

(19)

From Inequality (18) and (19), and Lemma 1 and 2, we have that

\[
\sum_{i:w_i \in S} p_i \leq m \delta \left( \sum_{i,k,w_i \in \mathcal{N}} \frac{c_i^1 + c_k^2 \epsilon}{c_i^1 + c_k^2 \epsilon} \right) \sum_{i:w_i \in S} \left( c_i^1 + c_i^2 \epsilon \right)
\]

\[
= m \delta \left( \sum_{i:w_i \in S} (c_i^1 + c_i^2 \epsilon) \right) \leq 2\delta \gamma H_m \text{C}_{OPT}
\]

Therefore, the proposed pSRC auction satisfies a \( 2\delta \gamma H_m \)-approximation to the optimal total payment.
### 4.3 Data Perturbation Mechanism

#### 4.3.1 Proposed Mechanism

As previously mentioned, any adversary curious about workers’ data could try to infer them utilizing the aggregated results if they are published directly. One example of such an adversary could be another competing platform hosting similar sensing tasks. The portion of workers’ data inferred with reasonable accuracy could be utilized by the adversary platform to calculate the results of its own tasks. In this way, it could reduce the number of workers recruited by itself, and thus its financial expense for worker recruiting.

To enable such inference, the adversary needs the information about workers’ weights, namely $\lambda_{i,j}$’s, defined in Equation (6). That is, it has to know $\alpha$ and $\theta$, which is usually feasible for the adversary platform. For similar sensing tasks, $\alpha$ is typically a common and standard design choice across different platforms, and workers’ skill levels for similar tasks tend to be similar as well. Therefore, $\theta$ can also be effectively estimated or inferred by the adversary platform using the methods mentioned in Section 3.2, such as utilizing workers’ sensory data about similar tasks collected during its past interactions with them as in [35, 36], using some of workers’ characteristics (e.g., reputation and experience for similar tasks) as in [37], and many others. To tackle such inference attack, we propose a novel data perturbation mechanism in Algorithm 4 by tailoring the Laplace mechanism in [22, 33] to our problem setting.

**Algorithm 4: Data Perturbation Mechanism**

**Input:** $(x_1, \ldots, x_N); \alpha, \beta$

**Output:** $(\tilde{x}_1, \ldots, \tilde{x}_N)$

1. \textbf{foreach} $j$ s.t. $\tau_j \in \mathcal{T}$ \textbf{do}
2. \hspace{1em} $n_j \leftarrow$ random noise sampled from Lap$(0, -\frac{\alpha_j}{\ln \beta_j})$ (line 2) and adds it to the aggregated result $x_j$ (line 3).
3. \hspace{1em} $	ilde{x}_j \leftarrow x_j + n_j$;
4. \hspace{1em} return $(\tilde{x}_1, \ldots, \tilde{x}_N)$;

Algorithm 4 takes as inputs the vector of the aggregated results $(x_1, \ldots, x_N)$ output by the data aggregation mechanism and the $\alpha$ and $\beta$ vector. $\alpha$ and $\beta$ are the same sets of platform-chosen parameters utilized by Algorithm 1, 2, and 3 with $\alpha_j, \beta_j \in (0, 1)$. For every task $\tau_j$, Algorithm 4 independently samples a random noise $n_j$ from the Laplacian distribution with mean 0 and scaling $-\frac{\alpha_j}{\ln \beta_j}$, denoted as Lap$(0, -\frac{\alpha_j}{\ln \beta_j})$ (line 2) and adds it to the aggregated result $x_j$ (line 3). Although adding Laplacian noise as in [22, 33] is a well-established approach to achieve differential privacy, the scaling of the Laplacian distribution is application specific and has to be carefully designed to achieve a desirable trade-off between privacy and data accuracy.

#### 4.3.2 Analysis

We firstly analyze the accuracy guarantee of Algorithm 4.

**Theorem 7.** The data perturbation mechanism given in Algorithm 4 satisfies

$$\Pr[|\tilde{X}_j - X_j| \geq \alpha_j] = \beta_j.$$  \hspace{1em} (20)

**Proof.** We use $N_j$ to denote the random variable representing the random noise sampled from the Laplacian distribution Lap$(0, -\frac{\alpha_j}{\ln \beta_j})$, i.e., $N_j \sim$ Lap$(0, -\frac{\alpha_j}{\ln \beta_j})$. Thus, $\Pr[|\tilde{X}_j - X_j| \geq \alpha_j] = \Pr[|N_j| \geq \alpha_j] = 2\Pr[N_j \geq \alpha_j] = 2\int_{\alpha_j}^{\infty} \frac{\ln \beta_j}{2\alpha_j} \exp \left(\frac{-z\ln \beta_j}{\alpha_j}\right) dz = \beta_j,$

which gives us $\Pr[|\tilde{X}_j - X_j| \geq \alpha_j] = \beta_j$.

Theorem 7 states that $(\alpha_j, \beta_j)$-accuracy is guaranteed for the perturbed result compared to the original one before perturbation for every task $\tau_j$. However, our ultimate goal is to achieve that the perturbed results has satisfactory accuracy compared to ground truths, which is proved in the following Theorem 8.

**Theorem 8.** For every task $\tau_j \in \mathcal{T}$, we have

$$\Pr[|\tilde{X}_j - x_j^*| \geq 2\alpha_j] \leq 1 - (1 - \beta_j)^2.$$  \hspace{1em} (21)

**Proof.** As discussed in Section 4.1 and 4.2, the aggregated result for every task $\tau_j$ satisfies that $\Pr[|\tilde{X}_j - x_j^*| \geq \alpha_j] \leq \beta_j$. From Theorem 7 and the fact that $X_j - x_j^*$ and $\tilde{X}_j - X_j$ are two independent random variables, we have

$$\Pr[|\tilde{X}_j - x_j^*| > 2\alpha_j] \leq \Pr[|\tilde{X}_j - X_j| + |X_j - x_j^*| > 2\alpha_j] \leq 1 - (1 - \beta_j)^2,$$

which gives us $\Pr[|\tilde{X}_j - x_j^*| \geq 2\alpha_j] \leq 1 - (1 - \beta_j)^2$.

Therefore, Theorem 8 gives us that $(2\alpha_j, 1 - (1 - \beta_j)^2)$-accuracy is satisfied for the perturbed result of every task $\tau_j$ compared to its ground truth. In Theorem 9, we analyze the privacy guarantee of the data perturbation mechanism.

**Theorem 9.** The data perturbation mechanism given in Algorithm 4 satisfies $\epsilon$-differential privacy, where the privacy budget $\epsilon = \max_{\tau_j \in \mathcal{T}} ( - \frac{\ln \beta_j}{\alpha_j})$.

**Proof.** For any $\mathcal{O} \subseteq \mathbb{R}$ and $r \in \mathbb{R}$, we use $\mathcal{O} - r$ to denote the set $\{x' = x - r | x \in \mathcal{O}\}$, and $x^{(i)}$ and $\tilde{x}^{(i)}$ to denote the aggregated result for task $\tau_j$ before and after perturbation when one entry $x_{i,j}$ changes. We have $|x_j - x_j^{(i)}| \leq 1$, and

$$\Pr[\tilde{X}_j \in O] = \Pr[N_j \in O - X_j] = \int_{z \in O - x_j} \frac{\ln \beta_j}{2\alpha_j} \exp \left(\frac{z\ln \beta_j}{\alpha_j}\right) dz \leq \exp \left(-\frac{\ln \beta_j}{\alpha_j}\right) \int_{z \in O - x_j^{(i)}} \frac{\ln \beta_j}{2\alpha_j} \exp \left(\frac{z\ln \beta_j}{\alpha_j}\right) dz \leq \exp \left(-\frac{\ln \beta_j}{\alpha_j}\right) \Pr[\tilde{X}_j^{(i)} \in \mathcal{O}].$$

Note that the previous analysis focuses on a specific task $\tau_j$. The overall privacy budget considering all tasks in $\mathcal{T}$ is thus $\epsilon = \max_{\tau_j \in \mathcal{T}} ( - \frac{\ln \beta_j}{\alpha_j})$.

### 4.4 Summary of Design Details

Thus far, we have finished the description of the design details of INCEPTION. Its incentive mechanism (Section 4.2) selects a set of winners that are more likely to provide reliable data and determines the payments to compensate their sensing and privacy costs. Meanwhile, it approximately minimizes the platform’s total payment (Theorem 6), and satisfies computational efficiency (Theorem 5), truthfulness (Theorem 3), and individual rationality (Theorem 4). Incorporating workers’ skill levels, the data aggregation mechanism (Section 4.1) provides aggregated results with high accuracy (Corollary 1), and the data perturbation mechanism (Section 4.3) adds carefully controlled noises to the aggregated results to achieve differential privacy (Theorem 9), and small degradation of data accuracy (Theorem 7).

Overall, INCEPTION guarantees $\max_{\tau_j \in \mathcal{T}} ( - \frac{\ln \beta_j}{\alpha_j})$-differential privacy and $(2\alpha_j, 1 - (1 - \beta_j)^2)$-accuracy for every task $\tau_j$ (Theorem 8). The platform could carefully select the parameter $\alpha_j, \beta_j \in (0, 1)$ for every task $\tau_j$ to achieve satisfactory guarantee for data accuracy and workers’ privacy.
5. PERFORMANCE EVALUATION

In this section, we introduce the baseline methods, and simulation settings, as well as results.

![Figure 2: Platform’s total payment (setting I)](image)

![Figure 3: Platform’s total payment (setting II)](image)

![Figure 4: Platform’s total payment (setting III)](image)

![Figure 5: Platform’s total payment (setting IV)](image)

![Figure 6: MAE of data after per-aggregation](image)

![Figure 7: EP after perturbation](image)

5.2 Simulation Settings

In our simulation, we generate \( x_{i,j} \) (i.e., worker \( u_i \)’s data about task \( \tau_j \)) from a normal distribution with mean \( \mu_{i,j} \) and standard deviation \( \sigma_{i,j} \) truncated within the range \([0, 1]\). The platform maintains the value of \( \theta_{i,j} \), calculated as:

\[
\theta_{i,j} = \frac{c_{i,j} \sigma_{i,j}}{\sqrt{2\pi}} \left( 2 \exp\left( \frac{-b_{i,j}^2}{2\sigma_{i,j}^2} \right) - \exp\left( \frac{-a_{i,j}^2}{2\sigma_{i,j}^2} \right) - \exp\left( \frac{-(1 - a_{i,j})^2}{2\sigma_{i,j}^2} \right) \right) + b_{i,j} \Phi\left( \frac{1 - a_{i,j}}{\sigma_{i,j}} \right) - \Phi\left( \frac{a_{i,j}}{\sigma_{i,j}} \right)
\]

where \( c_{i,j} = \Phi\left( \frac{-\mu_{i,j}}{\sigma_{i,j}} \right) - \Phi\left( -\frac{\mu_{i,j}}{\sigma_{i,j}} \right) \), \( b_{i,j} = \mu_{i,j} - x_{i,j}^*, \) and \( \Phi() \) denotes the c.d.f. of the standard normal distribution. We omit the derivation for \( \theta_{i,j} \) due to space limit. The parameter settings are given in Table 1.

### Table 1: Simulation settings

| Setting \( x, \beta, c_i, c_j, x_j^*, \mu_{i,j}, \sigma_{i,j}, |\Gamma| \) | N | K |
|---|---|---|
| I \[0, 0.1 \] | 1, 1, 15, 20 | 40 |
| II \[0, 0.1 \] | 1, 1, 15, 20 | 100 |
| III \[0, 0.1 \] | 1, 1, 25, 35 | 500 |
| IV \[0, 0.1 \] | 1, 1, 25, 35 | 1000 |

From Table 1, we observe that the VCG auction has excessively long running time so that it can hardly be utilized in practice. The running time of the VCG auction lower bounds that of the auction that gives us close-to-optimal total payment. Next, we compare the execution time of the VCG and the BPE-Greedy auction.

5.3 Simulation Results

Figure 2 and 3 show that the platform’s total payment of the pSRC auction is far less than that of the BPE-Greedy auction and fairly close to the optimal social cost given by the VCG auction. Since the optimal social cost lower bounds the optimal total payment, the pSRC auction thus gives us close-to-optimal total payment. Next, we compare the execution time of the VCG and the BPE-Greedy auction.

From Table 2, we observe that the VCG auction has excessively long running time so that it can hardly be utilized in practice. The running time of the VCG auction lower bounds that of the auction that gives us the optimal total payment, because solving the pSRC-SCM problem is in fact easier and faster than solving the pSRC-TPM problem. Hence, calculating the optimal total payment becomes computationally infeasible in practice. However, the execution time of the pSRC auction keeps in the order of microsecond, which is much less that of the VCG auction.

### Table 2: Execution time (s) for setting I and II

<table>
<thead>
<tr>
<th>N</th>
<th>91</th>
<th>95</th>
<th>99</th>
<th>103</th>
<th>107</th>
<th>111</th>
<th>115</th>
<th>119</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCG</td>
<td>20.23</td>
<td>20.11</td>
<td>20.23</td>
<td>20.37</td>
<td>20.85</td>
<td>8.66</td>
<td>8.79</td>
<td>10.0</td>
</tr>
<tr>
<td>pSRC</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>K</td>
<td>21</td>
<td>25</td>
<td>29</td>
<td>33</td>
<td>37</td>
<td>41</td>
<td>45</td>
<td>49</td>
</tr>
<tr>
<td>VCG</td>
<td>0.300</td>
<td>0.676</td>
<td>13.09</td>
<td>30.60</td>
<td>1063</td>
<td>1169</td>
<td>1350</td>
<td>1677</td>
</tr>
<tr>
<td>pSRC</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

In Figure 4 and 5, we show our simulation results about the platform’s total payment for setting III and IV with larger-size problem instances where the VCG auction is not able to terminate in reasonable time. We can observe that the proposed pSRC auction still gives us a total payment far

5.1 Baseline Methods

Ideally we need to compare the proposed pSRC auction with a truthful and individual rational auction that returns exact optimal solutions to the pSRC-TPM problem. However, because solving the pSRC-TPM problem is notoriously challenging, we instead use the following VCG auction [39, 40] as one of the baseline methods. The VCG auction solves the pSRC-SCM problem optimally and pays every winner according to the VCG payment. This choice is reasonable as the optimal social cost offers a lower bound to the optimal total payment as proved in Lemma 1. Hence, a good approximation to the optimal social cost indicates a better approximation to the optimal total payment.

Another baseline method is the bidding price effectiveness greedy (BPE-Greedy) auction. Initially, it sorts workers according to their bidding price effectiveness. Winners are selected in this order until the feasibility of the pSRC-TPM problem is satisfied. Its pricing mechanism pays every winner her critical payment as Algorithm 3 does. It is easily provable that the BPE-Greedy auction also satisfies truthfulness and individual rationality.

Furthermore, we compare our weighted data aggregation mechanism with two other schemes that calculate the mean and median of winners’ data, respectively.
We evaluate the accuracy guarantee of INCEPTION in setting II with a minor change of the parameter $\beta_j$, i.e., $\beta_j$ is fixed as 0.05 for every task $\tau_j$ to simplify presentation. We compare the mean absolute error (MAE) for all tasks, defined as $\text{MAE} = \frac{1}{K} \sum_{j=1}^{K} \sum_{\tau \in \mathcal{T}} |x_j - \hat{x}_j|$, of the weighted aggregation mechanism given in Algorithm 1 with those of the mean and median aggregation. The simulation for each combination of worker and task number is repeated for 10000 times and the means and standard deviations of the MAEs are plotted. We observe from Figure 6 that the MAE of our weighted aggregation is far less than those of the mean and median aggregation. Then, we show simulation results regarding $\text{Pr}[|\hat{x}_j - x_j| \geq \alpha_j]$, referred to as the error probability (EP) of the perturbed results for task $\tau_j$. After 10000 repetitions of the simulation for any specific combination of worker and task number, empirical values for the EPs are calculated and we plot the means and standard deviations of the empirical EPs over all tasks. From Figure 7, we observe that the empirical EPs are far less than the required upper bound (i.e., $1 - (1 - \beta_j)^2 = 1 - (1 - 0.05)^2 = 0.0975$).

6. CONCLUSION

We propose INCEPTION, a novel MCS system framework that integrates an incentive, a data aggregation, and a data perturbation mechanism. Its incentive mechanism selects reliable workers, and compensates their costs for sensing and privacy leakage, which meanwhile satisfies truthfulness and individual rationality. Its data aggregation mechanism incorporates workers’ reliability to generate highly accurate aggregated results, and its data perturbation mechanism ensures satisfactory guarantee for workers’ privacy, as well as the accuracy for the final perturbed results. The desirable properties of INCEPTION are validated through both theoretical analysis and extensive simulations.

References