Incentive Mechanism for Privacy-Aware Data Aggregation in Mobile Crowd Sensing Systems

Haiming Jin, *Member, IEEE*, Lu Su, *Member, IEEE, ACM*, Houping Xiao, *Student Member, IEEE*, and Klara Nahrstedt, *Fellow, IEEE, ACM*

***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 the 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.**

***Index Terms*— Incentive mechanism, data aggregation, privacy preservation, mobile crowd sensing.**

I. 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]–[4] 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 [5]–[25], we capture such interactive effect, and propose INCEPTION,[1] 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, environment 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 aggregated results closer to the data provided by more reliable workers. Therefore, we propose a weighted data aggregation mechanism that incorporates workers' diverse *reliability* 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 adver-

Manuscript received March 23, 2017; revised December 12, 2017; accepted April 25, 2018; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Shin. Date of publication August 16, 2018; date of current version October 15, 2018. This work was supported in part by the National Science Foundation under Grants CNS-1330491 and 1652503 and in part by the Ralph and Catherine Fisher Grant. *(Corresponding author: Haiming Jin.)*

H. Jin was with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA. He is now with the John Hopcroft Center for Computer Science and the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jinhaiming@sjtu.edu.cn).

L. Su and H. Xiao are with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: lusu@buffalo.edu; houpingx@buffalo.edu).

K. Nahrstedt is with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA, and also with the Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: klara@illinois.edu).

Digital Object Identifier 10.1109/TNET.2018.2840098

[1]The name INCEPTION comes from INCEtive, Privacy, and data aggregaTION.

saries 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.

II. Related Work

Game theory has been widely adopted, thus far, by the research community in the design of incentive mechanisms for MCS systems [5]–[25] so as to tackle workers' strategic behaviors. Specifically, these prior work utilize either auction [13]–[25] or other game-theoretic models [5]–[12]. Although with different objectives, including maximizing social welfare [12]–[17] or platform's profit [5]–[9], [18]–[22], and minimizing social cost [23] or platform's payment [10], [11], [24], [25], 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 [26] and [27]) 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 [26] and [27] is not necessarily obtained by sensing, and thus, sensing costs are not considered by [26] and [27].

One line of past literature [28]–[33], 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 [28]–[32]. Furthermore, unlike this paper, most of these work do not consider the issue of providing incentives to workers. Another set of existing work [34]–[37], 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 [34], [35] or bid privacy [36], [37] within the incentive mechanisms.

III. Preliminaries

In this section, we give an overview of INCEPTION, and describe the task model, reliability level model, auction model, as well as design objectives.

A. System Overview

INCEPTION is an MCS system framework consisting of a cloud-based platform and a set of N participating workers, denoted as $\mathcal{N} = \{w_1, \cdots, w_N\}$. The platform hosts a set of K sensing tasks, denoted as $\mathcal{T} = \{\tau_1, \cdots, \tau_K\}$, where each task $\tau_j \in \mathcal{T}$ requires workers to locally sense a specific object or phenomenon, and report the sensory data to the platform. If worker w_i is selected to execute task τ_j, she will provide her data $x_{i,j}$ to the platform. We define $\mathbf{x} = [x_{i,j}] \in \big(\mathcal{X} \cup \{\perp\}\big)^{N \times K}$ as the matrix containing all workers' data, where $\mathcal{X}$ denotes the range of tasks' sensory data, and $x_{i,j} = \perp$ means that task τ_j is not executed by worker w_i. To cancel out the errors from individual workers, for each task τ_j, the platform aggregates workers' data into an aggregated result, denoted as x_j, which is used as an estimate of the task's ground truth value x_j^*, unknown to the platform and the workers.

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 ϵ-differential privacy defined in Definition 1.

Definition 1 (Differential Privacy): We denote $M : (\mathcal{X} \cup \{\perp\})^{N \times K} \rightarrow \mathbb{R}^{K \times 1}$ as a mechanism that maps any input data matrix to a perturbed result vector. Then, the mechanism M is ϵ-differentially private if and only if for any two data matrices $\mathbf{x}$ and $\mathbf{x}'$ that differ in only one entry and any $\mathcal{A} \subseteq \mathbb{R}^{K \times 1}$, we have

$$\Pr[M(\mathbf{x}) \in \mathcal{A}] \leq \exp(\epsilon) \Pr[M(\mathbf{x}') \in \mathcal{A}], \tag{1}$$

where ϵ 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.

- Firstly, the platform announces the set of sensing tasks $\mathcal{T}$ and an upper bound of the privacy budget ϵ, such as $\epsilon \leq 0.5$, to workers (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 to the platform her bid $b_i = (\Gamma_i, b_i^s, b_i^p)$ which is a triple containing the set of sensing tasks Γ_i she wants

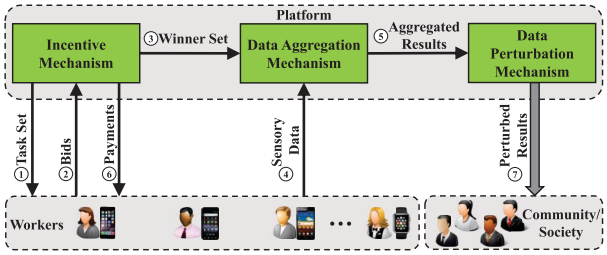


Fig. 1. Framework of INCEPTION (where circled numbers represent the order of the events).

to execute, as well as her bidding prices for executing them b_i^s and unit privacy loss b_i^p (step ②). Based on workers' bids, the platform determines the set of winners $\mathcal{S} \subseteq \mathcal{N}$ and the payment p_i to every winner w_i (step ③). Losers of the auction do not execute tasks and receive no payments. We denote workers' bid and payment profile as $\mathbf{b} = (b_1, \cdots, b_N)$ and $\mathbf{p} = (p_1, \cdots, p_N)$, respectively.

- *Data Aggregation Mechanism:* Next, the platform collects winners' sensory data (step ④) and calculates an aggregated result x_j for each task τ_j (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 ϵ (step ⑥), such as $\epsilon = 0.25$. The design rationale for keeping the exact value of ϵ confidential to workers at the bidding stage and revealing it together with the payments is described in detail in Section IV-B.3.
- *Data Perturbation Mechanism:* Finally, the platform adds random noises to the original aggregated results and publishes the perturbed ones (step ⑦). We use $\widehat{x}_j$ to denote the perturbed result for task τ_j.

B. Task Model

In this paper, we target a general scenario, where our MCS system collects *heterogeneous* types of sensory data from participating workers. That is, some of the tasks held by the platform (e.g., environmental monitoring) require workers to submit *continuous data* (e.g., temperature, humidity), whereas others (e.g., geotagging) collect *categorical data* (e.g., whether or not potholes exist on a specific road segment). In the rest of this paper, we refer to the former as *continuous tasks*, and the latter as *categorical tasks*. Furthermore, we use $\mathcal{T}_{\text{con}}$ and $\mathcal{T}_{\text{cat}}$ to denote the set of continuous tasks and categorical tasks, respectively. Obviously, $\mathcal{T} = \mathcal{T}_{\text{con}} \cup \mathcal{T}_{\text{cat}}$.

To avoid unnecessary complications to the analysis and presentation of our results, we assume that, for each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, the ground truth x_j^* and any worker w_i's data $x_{i,j}$ are within the range $[0, 1]$. Such assumption is without loss of generality, as we could convert data of arbitrary value to be in the range $[0, 1]$ by proper normalization. For the same reason, we assume that all categorical tasks in $\mathcal{T}_{\text{cat}}$ are binary classification tasks with ground truths x_j^*'s, as well as workers' labels, taking values from the set $\{+1, -1\}$. Note that this is also a mild assumption, as for any binary classification task, we could assign the label $+1$ to one class, and -1 to the other.

C. Reliability Level Model

Before task τ_j 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 *reliability level* for continuous and categorical tasks, respectively, in Definition 2 and 3.

Definition 2 (Reliability Level for Continuous Task): Worker w_i's reliability level $\theta_{i,j}$ for a continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ is defined as the expected absolute difference between her data and the ground truth, i.e.,

$$\theta_{i,j} = \mathbb{E}[|X_{i,j} - x_j^*|] \in [0, 1], \tag{2}$$

where the expectation is taken over the randomness of $X_{i,j}$.

Definition 3 (Reliability Level for Categorical Task): Worker w_i's reliability level $\theta_{i,j}$ for a categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$ is defined as the probability that she provides a correct label about this task, i.e.,

$$\theta_{i,j} = \Pr[X_{i,j} = x_j^*] \in [0, 1]. \tag{3}$$

We use $\boldsymbol{\theta} = [\theta_{i,j}] \in [0, 1]^{N \times K}$ to denote the reliability level matrix of all workers. We assume that the reliability level matrix $\boldsymbol{\theta}$ is *a priori* known to the platform. In practice, the platform can keep a historical record of $\boldsymbol{\theta}$, which can be obtained by many methods. For example, since a worker's reliability 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 reliability levels for similar tasks as in [38]. In scenarios where ground truths are not available, $\boldsymbol{\theta}$ can still be effectively estimated utilizing workers' previously submitted sensory data about similar tasks by algorithms proposed in [39] and [40] 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 [41].

D. 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 $\mathcal{T}$ 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 [42], 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)*, defined in Definition 4, as the incentive mechanism.

Definition 4 (pSRC Auction): In a single-minded reverse combinatorial auction with privacy cost (pSRC auction), each worker w_i has only one interested bundle Γ_i^*. Her cost of executing the bundle of tasks, namely sensing cost, is denoted as c_i^s (unknown to the platform). Additionally, she has a cost for privacy leakage, namely privacy cost, denoted as $C_i^p(\epsilon)$, if ϵ-differential privacy is guaranteed. Hence, worker w_i's cost function is defined as in Equation (4).

$$C_i(\Gamma, \epsilon) = \begin{cases} c_i^s + C_i^p(\epsilon), & \text{if } \Gamma \subseteq \Gamma_i^* \\ +\infty, & \text{otherwise.} \end{cases} \tag{4}$$

For the tasks that do not belong to worker w_i's interested bundle Γ_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 (4).

A major difference between the cost function defined in Equation (4) and those in [5]–[25] is that the privacy cost

$C_i^p(\epsilon)$ 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 $C_i^p(\epsilon)$ is positively correlated with the privacy budget ϵ, because ϵ in fact captures the amount of privacy leakage of the MCS system. Therefore, we adopt the natural linear model for privacy cost as in [26] and [27] where $C_i^p(\epsilon) = c_i^p \epsilon$ with c_i^p representing worker w_i's cost for unit privacy leakage. Similar to c_i^s, c_i^p is also unknown to the platform. Next, we define a worker's utility in Definition 5.

Definition 5 (Worker's Utility): Any worker w_i's utility u_i is defined as

$$u_i = \begin{cases} p_i - c_i^s - c_i^p \epsilon, & \text{if } w_i \in \mathcal{S} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Apart from workers' utilities, we are also interested in the platform's total payment defined in Definition 6.

Definition 6 (Platform's Total Payment): Given the payment profile $\mathbf{p}$ and the winner set $\mathcal{S}$, the platform's total payment is $P = \sum_{i:w_i \in \mathcal{S}} p_i$.

E. Design Objective

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

Since workers are strategic in our model, it is possible that a worker w_i submits a bid (Γ_i, b_i^s, b_i^p) that deviates from the true value $(\Gamma_i^*, c_i^s, c_i^p)$. However, one of our objectives is to design a *truthful* incentive mechanism defined in Definition 7.

Definition 7 (Truthfulness): A pSRC auction is truthful if and only if bidding the true value $(\Gamma_i^*, c_i^s, c_i^p)$ is the dominant strategy for each worker w_i, i.e., bidding $(\Gamma_i^*, c_i^s, c_i^p)$ maximizes each worker w_i's utility for all possible values of other workers' bids and the privacy budget ϵ.

By Definition 7, we aim to ensure the truthful bidding of the interested bundle Γ_i^*, the sensing cost c_i^s, and the cost for unit privacy leakage c_i^p for every worker w_i. Apart from truthfulness, another desirable and necessary property is *individual rationality* defined in Definition 8.

Definition 8 (Individual Rationality): A pSRC auction is individual rational if and only if no worker receives negative utility, i.e., we have $u_i \geq 0$ 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 III-A, we aim to design an MCS system that ensures *ϵ-differential privacy*. However, the perturbation added to the aggregated results inevitably impairs their accuracy. Next, we formally define the concept of (α, β)-accuracy for continuous tasks in Definition 9.

Definition 9 ((α, β)-Accuracy): For two random variables Y_1 and Y_2 within the range $[0, 1]$, Y_1 is (α, β)-accurate to Y_2, if and only if $\Pr[|Y_1 - Y_2| \geq \alpha] \leq \beta$, where $\alpha, \beta \in (0, 1)$. Note that Y_2 could also be a constant.

We use $\widehat{X}_j$ to denote the random variable corresponding to $\widehat{x}_j$ (i.e., the perturbed result for task τ_j). Facing the trade-off between privacy and accuracy, we need to carefully control the amount of noises added to the aggregated results and ensure that, for each continuous task τ_j, $\widehat{X}_j$ is (α, β)-accurate to the ground truth x_j^* with sufficiently small α and β within $(0, 1)$. That is, we aim to ensure that the perturbed results of all continuous tasks are fairly close to the ground truths with high probability. For categorical tasks, we adopt the notion of γ-accuracy, which is formally defined in Definition 10.

Definition 10 (γ-Accuracy): For two random variables Z_1 and Z_2 that take values from the set $\{+1, -1\}$, Z_1 is γ-accurate to Z_2, if and only if $\Pr[Z_1 \neq Z_2] \leq \gamma$, where $\gamma \in (0, 1)$. Note that Z_2 could also be a constant.

For each categorical task τ_j, we aim to ensure that the perturbed result $\widehat{X}_j$ is γ-accurate to the ground truth x_j^* with a sufficiently small $\gamma \in (0, 1)$, which means that the perturbed results of all categorical tasks are equal to the 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.

IV. Design Details

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

A. Data Aggregation Mechanism

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 the 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 τ_j based on workers' data. That is, given the winner set $\mathcal{S}$ determined by the incentive mechanism, the aggregated result x_j of each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ is calculated as

$$x_j = \sum_{i:w_i \in \mathcal{S}, \tau_j \in \Gamma_i} \lambda_{i,j} x_{i,j}, \tag{6}$$

where $\lambda_{i,j} > 0$ is the weight of worker w_i on this task with $\sum_{i:w_i \in \mathcal{S}, \tau_j \in \Gamma_i} \lambda_{i,j} = 1$ for every continuous task τ_j. Similarly, for each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, we calculate the aggregated result x_j as

$$x_j = \text{sign}\left(\sum_{i:w_i \in \mathcal{S}, \tau_j \in \Gamma_i} \lambda_{i,j} x_{i,j} \right), \tag{7}$$

where, similar to continuous tasks, $\lambda_{i,j} > 0$ is worker w_i's weight on this task $\sum_{i:w_i \in \mathcal{S}, \tau_j \in \Gamma_i} \lambda_{i,j} = 1$. Furthermore, function $\text{sign}(z)$ equals to $+1$, when $z \geq 0$, and -1 otherwise.

The motivation for utilizing weighted aggregation is to capture the effect of workers' diverse reliability 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 [39], [40] utilize such weighted aggregation to calculate the aggregated results. Since the accuracy of the aggregated results highly depends on how the weight $\lambda_{i,j}$'s are chosen, we propose the data aggregation mechanism in Algorithm 1.

Algorithm 1 Data Aggregation Mechanism

Input: $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\mathbf{b}$, $\mathbf{x}$, $\mathcal{S}$, $\mathcal{T}$;
Output: $(x_1, \cdots, x_K)$;
1 **foreach** j *s.t.* $\tau_j \in \mathcal{T}$ **do**
2 **if** $\tau_j \in \mathcal{T}_{\text{con}}$ **then**
   // Calculate the aggregated result of a continuous task
3 $x_j \leftarrow \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i} \frac{(\alpha_j-\theta_{i,j})x_{i,j}}{\sum_{k:w_k\in\mathcal{S},\tau_j\in\Gamma_k}(\alpha_j-\theta_{k,j})}$;
4 **else**
   // Calculate the aggregated result of a categorical task
5 $x_j \leftarrow \text{sign}\big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i} \frac{(2\theta_{i,j}-1)x_{i,j}}{\sum_{k:w_k\in\mathcal{S},\tau_j\in\Gamma_k}(2\theta_{k,j}-1)}\big)$;
6 **return** $(x_1, \cdots, x_K)$;

Besides the reliability level matrix $\boldsymbol{\theta}$, the bid profile $\mathbf{b}$, workers' data $\mathbf{x}$, the winner set $\mathcal{S}$, as well as the task set $\mathcal{T}$, Algorithm 1 also takes as input a vector of positive real numbers $\boldsymbol{\alpha}$, where each element α_j corresponds to one continuous task τ_j. These α_j's are parameters chosen by the platform, such that $\max_{i:\tau_j\in\Gamma_i}\theta_{i,j} < \alpha_j < 0.5$. Note that, for a continuous task τ_j, large $\theta_{i,j}$ indicates low reliability, and any worker w_i with $\theta_{i,j} \geq 0.5$ will not be selected by the incentive mechanism. The aggregated result x_j of every continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ is calculated (line 3) using Equation (6) with the weight

$$\lambda_{i,j} = \frac{\alpha_j - \theta_{i,j}}{\sum_{k:w_k\in\mathcal{S},\tau_j\in\Gamma_k}(\alpha_j - \theta_{k,j})}, \quad \forall w_i \in \mathcal{S},\ \tau_j \in \Gamma_i. \quad (8)$$

By Equation (8), worker w_i's weight for a continuous task τ_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.

For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, we calculate its aggregated result x_j (line 5) using Equation (7) with the weight,

$$\lambda_{i,j} = \frac{2\theta_{i,j} - 1}{\sum_{k:w_k\in\mathcal{S},\tau_j\in\Gamma_k}(2\theta_{k,j} - 1)}, \quad \forall w_i \in \mathcal{S},\ \tau_j \in \Gamma_i. \quad (9)$$

Note that large $\theta_{i,j}$ for a categorical task implies high reliability, and the incentive mechanism will not select any worker w_i with $\theta_{i,j} \leq 0.5$ to execute this task. Following a similar philosophy as calculating the aggregated result of a continuous task, the data from workers with higher reliability are counted more in the calculation of a categorical task's aggregated result, as well. Formal analysis about the data aggregation mechanism is provided in Section IV-A.2.

2) Analysis: In this section, we first analyze Algorithm 1's guarantee of aggregation accuracy for continuous tasks. In the following Lemma 1, we establish an upper bound for the accuracy of the aggregated result x_j of each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ compared to its ground truth x_j^*. In the rest of our analyses, we use X_j to denote the random variable representing any task τ_j's aggregated result x_j.

Lemma 1: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, given the winner set $\mathcal{S}$, the reliability level matrix $\boldsymbol{\theta}$, the vector of platform-chosen parameter $\boldsymbol{\alpha}$, as well as workers' weights $\lambda_{i,j}$'s on this task, we have that

$$\Pr\big[\big|X_j - x_j^*\big| \geq \alpha_j\big] \leq \exp\left(-\frac{2\big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}(\alpha_j-\theta_{i,j})\big)^2}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\right) \quad (10)$$

by aggregating workers' data according to Equation (6).

Proof: From Equation (6), for each continuous task τ_j, we have that

$$\begin{aligned}\big|X_j - x_j^*\big| &= \Big|\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}X_{i,j} - x_j^*\Big| \\ &= \Big|\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}\big(X_{i,j} - x_j^*\big)\Big| \\ &\leq \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\Big|\lambda_{i,j}\big(X_{i,j} - x_j^*\big)\Big|.\end{aligned}$$

We define a random variable L_j for every continuous task τ_j as $L_j = \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\big|\lambda_{i,j}(X_{i,j} - x_j^*)\big|$, which is the sum of random variables $L_{i,j}$'s with $L_{i,j} = |\lambda_{i,j}(X_{i,j} - x_j^*)| \in [0, \lambda_{i,j}]$. Thus,

$$\mathbb{E}[L_j] = \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}\mathbb{E}\big[\big|X_{i,j} - x_j^*\big|\big] = \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}\theta_{i,j}.$$

Therefore, from the Hoeffding bound, we have

$$\begin{aligned}&\Pr\big[\big|X_j - x_j^*\big| \geq \alpha_j\big] \\ &\leq \Pr\Big[\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\big|\lambda_{i,j}\big(X_{i,j} - x_j^*\big)\big| \geq \alpha_j\Big] = \Pr[Y_j \geq \alpha_j] \\ &\leq \exp\left(-\frac{2\big(\alpha_j - \mathbb{E}[Y_j]\big)^2}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\right) \\ &= \exp\left(-\frac{2\big(\alpha_j - \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}\theta_{i,j}\big)^2}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\right) \\ &= \exp\left(-\frac{2\big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}(\alpha_j-\theta_{i,j})\big)^2}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\right),\end{aligned}$$

which exactly proves this lemma. □

Clearly, Lemma 1 gives us an upper bound for the probability $\Pr[|X_j - x_j^*| \geq \alpha_j]$ for each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$. Then, in the following Theorem 1, we will prove that this upper bound is minimized by our proposed Algorithm 1.

Theorem 1: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, the data aggregation mechanism proposed in Algorithm 1 minimizes the upper bound of the probability $\Pr[|X_j - x_j^*| \geq \alpha_j]$ established in Lemma 1, and ensures that

$$\Pr\big[\big|X_j - x_j^*\big| \geq \alpha_j\big] \leq \exp\left(-2\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j - \theta_{i,j})^2\right). \quad (11)$$

Proof: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, we denote $\boldsymbol{\lambda}_j = [\lambda_{i,j}]$ as the vector that contains every $\lambda_{i,j}$ such that $w_i \in \mathcal{S}$ and $\tau_j \in \Gamma_i$. Therefore, minimizing the upper bound

of $\Pr[|X_j - x_j^*| \geq \alpha_j]$ established in Lemma 1 is equivalent to maximizing the function $\varphi(\boldsymbol{\lambda}_j)$ defined as

$$\varphi(\boldsymbol{\lambda}_j) = \frac{\Big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}(\alpha_j-\theta_{i,j})\Big)^2}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}.$$

From the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\varphi(\boldsymbol{\lambda}_j) &\leq \frac{\Big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2\Big)\Big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j-\theta_{i,j})^2\Big)}{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\\ &= \sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j-\theta_{i,j})^2\end{aligned}$$

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

Using the fact that $\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j} = 1$, we have

$$\lambda_{i,j} = \frac{\alpha_j-\theta_{i,j}}{\sum_{k:w_k\in\mathcal{S},\tau_j\in\Gamma_k}(\alpha_j-\theta_{k,j})}. \tag{12}$$

Therefore, when $\lambda_{i,j}$'s satisfy Equation (12), we have

$$\Pr\big[\big|X_j - x_j^*\big| \geq \alpha_j\big] \leq \exp\left(-2\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j-\theta_{i,j})^2\right),$$

which is exactly the Equation (11) in Theorem 1. □

By Theorem 1, for each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, the data aggregation mechanism proposed in Algorithm 1 upper bounds the probability of $\Pr[|X_j - x_j^*| \geq \alpha_j]$ by $\exp\big(-2\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j-\theta_{i,j})^2\big)$ which is the minimum value of the upper bound established in Lemma 1 for this probability. Then, we introduce Corollary 1 which is directly utilized in the design of the incentive mechanism in Section IV-B.

Corollary 1: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, if

$$\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\big(\alpha_j-\theta_{i,j}\big)^2 \geq \frac{1}{2}\ln\left(\frac{1}{\beta_j}\right), \tag{13}$$

then the data aggregation mechanism proposed in Algorithm 1 ensures that $\Pr[|X_j - x_j^*| \geq \alpha_j] \leq \beta_j$, where $\beta_j \in (0,1)$ is a parameter chosen by the platform for this task. We use $\boldsymbol{\beta}$ to denote the vector, where each element β_j corresponds to one continuous task τ_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

$$\exp\left(-2\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(\alpha_j-\theta_{i,j})^2\right) \leq \beta_j,$$

which is equivalent to exactly

$$\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\big(\alpha_j-\theta_{i,j}\big)^2 \geq \frac{1}{2}\ln\left(\frac{1}{\beta_j}\right). \tag{14}$$

Therefore, together with Theorem 1, we know that Inequality (14) implies $\Pr[|X_j - x_j^*| \geq \alpha_j] \leq \beta_j$. □

Corollary 1 states that (α_j, β_j)-accuracy is guaranteed for the aggregated result of each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ compared to its ground truth x_j^*, if the condition specified by Inequality (13) is satisfied by the set of selected winners $\mathcal{S}$ in the incentive mechanism proposed in Section IV-B.

Next, we introduce the analyses on Algorithm 1's aggregation accuracy for categorical tasks in the following Lemma 2, Theorem 2, and Corollary 2, which are adapted from [14, Th. 1, Corollary 1].

Lemma 2: For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, given the winner set $\mathcal{S}$, the reliability level matrix $\boldsymbol{\theta}$, as well as workers' weights $\lambda_{i,j}$'s on this task, we have that

$$\Pr[X_j \neq x_j^*] \leq \exp\left(-\frac{\big(\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}(2\theta_{i,j}-1)\big)^2}{2\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}\lambda_{i,j}^2}\right) \tag{15}$$

by aggregating workers' data according to Equation (7).

Theorem 2: For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, the data aggregation mechanism proposed in Algorithm 1 minimizes the upper bound of the probability $\Pr[X_j \neq x_j^*]$ established in Lemma 2, and ensures that

$$\Pr[X_j \neq x_j^*] \leq \exp\left(-\frac{\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(2\theta_{i,j}-1)^2}{2}\right). \tag{16}$$

Corollary 2: For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, if

$$\sum_{i:w_i\in\mathcal{S},\tau_j\in\Gamma_i}(2\theta_{i,j}-1)^2 \geq 2\ln\left(\frac{1}{\gamma_j}\right), \tag{17}$$

then the data aggregation mechanism proposed in Algorithm 1 ensures that $\Pr[X_j \neq x_j^*] \leq \gamma_j$, where $\gamma_j \in (0,1)$ is a parameter chosen by the platform for this task. We use $\boldsymbol{\gamma}$ to denote the vector, where each element γ_j corresponds to one categorical task τ_j.

The proofs of Lemma 2, Theorem 2, and Corollary 2 are omitted in this paper, because they can be adapted from those of [14, Th. 1, Corollary 1] with minor changes. Clearly, they are counterparts of Lemma 1, Theorem 1, and Corollary 1 for categorical tasks, and collectively ensure that γ_j-accuracy is guaranteed for the aggregated result of each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, as long as Inequality (17) is satisfied by the winners selected by the incentive mechanism. Next, in Section IV-B, we introduce the design of INCEPTION's incentive mechanism, which is based on the data aggregation mechanism proposed in Algorithm 1.

B. Incentive Mechanism

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

1) Mathematical Formulation: As mentioned in Section III-D, our incentive mechanism is based on the pSRC auction defined in Definition 4. 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:

$$\min \sum_{i:w_i\in\mathcal{N}} p_i y_i \tag{18}$$

$$\text{s.t.} \sum_{i:w_i\in\mathcal{N},\tau_j\in\Gamma_i}\big(\alpha_j-\theta_{i,j}\big)^2\, y_i \geq \frac{1}{2}\ln\left(\frac{1}{\beta_j}\right), \quad \forall \tau_j \in \mathcal{T}_{\text{con}} \tag{19}$$

$$\sum_{i:w_i\in\mathcal{N},\tau_j\in\Gamma_i}(2\theta_{i,j}-1)^2\ y_i \geq 2\ln\left(\frac{1}{\gamma_j}\right),\quad \forall\tau_j\in\mathcal{T}_{\text{cat}} \tag{20}$$

$$y_i\in\{0,1\},\ p_i\in[0,+\infty),\quad \forall w_i\in\mathcal{N} \tag{21}$$

Constants: The pSRC-TPM problem takes as inputs the worker set $\mathcal{N}$, the continuous and categorical task set $\mathcal{T}_{\text{con}}$ and $\mathcal{T}_{\text{cat}}$, workers' bid profile $\mathbf{b}$, the reliability level matrix $\boldsymbol{\theta}$, and the $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ vector.

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

Objective function: The objective function given by $\sum_{i:w_i\in\mathcal{N}}p_iy_i=\sum_{i:w_i\in\mathcal{S}}p_i$ is exactly the total payment made by the platform to all winners.

Constraints: Constraint (19) is equivalent to Inequality (13) given in Corollary 1, which specifies the condition that the selected winners should satisfy. By Corollary 1, any feasible solution $\mathbf{y}$ to the pSRC-TPM problem gives a winner set $\mathcal{S}$ which ensures that the aggregated result of every continuous task $\tau_j\in\mathcal{T}_{\text{con}}$ is (α_j,β_j)-accurate to the ground truth x_j^*. Similarly, Constraint (20) is equivalent to Inequality (7), which ensures that γ_j-accuracy for each categorical task $\tau_j\in\mathcal{T}_{\text{cat}}$ is satisfied by the winner set $\mathcal{S}$ given by any feasible solution $\mathbf{y}$ to the pSRC-TPM problem. To simplify presentation, we introduce the following extra notations. For each worker $w_i\in\mathcal{N}$ and task $\tau_j\in\mathcal{T}$, we define

$$q_{i,j}=\begin{cases}(\alpha_j-\theta_{i,j})^2, & \text{if } \tau_j\in\mathcal{T}_{\text{con}}\\ (2\theta_{i,j}-1)^2, & \text{if } \tau_j\in\mathcal{T}_{\text{cat}},\end{cases} \tag{22}$$

and for each task $\tau_j\in\mathcal{T}$, we define

$$Q_j=\begin{cases}\frac{1}{2}\ln\left(\frac{1}{\beta_j}\right), & \text{if } \tau_j\in\mathcal{T}_{\text{con}}\\ 2\ln\left(\frac{1}{\gamma_j}\right), & \text{if } \tau_j\in\mathcal{T}_{\text{cat}}.\end{cases} \tag{23}$$

Furthermore, we define $\mathbf{q}=[q_{i,j}]\in[0,+\infty)^{N\times K}$ and $\mathbf{Q}=[Q_j]\in[0,+\infty)^{K\times 1}$. Therefore, Constraint (19) and (20) can be simplified and merged into the following Constraint (24).

$$\sum_{i:w_i\in\mathcal{N},\tau_j\in\Gamma_i}q_{i,j}y_i\geq Q_j,\quad\forall\tau_j\in\mathcal{T}. \tag{24}$$

Besides Constraint (19) and (20), any feasible solution to the pSRC-TPM problem should also satisfy two other inherent constraints, namely truthfulness and individual rationality. These two implicit constraints impose additional restrictions to p_i's besides non-negativity. Because of the difficulty in mathematically formulating the two constraints, we take them into consideration without explicitly formulating them, in the pSRC-TPM problem.

In Theorem 3, we prove the NP-hardness of the pSRC-TPM problem.

Theorem 3: The pSRC-TPM problem is NP-hard.

Proof: We consider a special case of the pSRC-TPM problem with a constant payment profile $\mathbf{p}$ and the truthfulness and individual rationality constraints relaxed. 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 $\mathcal{T}=\{\tau_1,\cdots,\tau_K\}$ and a set of subsets of $\mathcal{T}$ defined as $\mathcal{R}=\{\Gamma_1,\cdots,\Gamma_N\}$. Each set $\Gamma_i\in\mathcal{R}$ has a non-negative weight p_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 Γ_i to Γ_i' where each element $\tau_j\in\Gamma_i$ has $a_{i,j}\in\mathbb{Z}^+$ copies and require each τ_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. □

2) Proposed Mechanism: Because of the NP-hardness of the pSRC-TPM problem proved in Theorem 3, 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 design rationale of Algorithm 2 is that we seek to ensure the *monotonicity* property, i.e., if a worker w_i wins the auction by bidding (Γ_i,b_i^s,b_i^p), she will also win by bidding $(\widetilde{\Gamma}_i,\widetilde{b}_i^s,\widetilde{b}_i^p)$ with $\widetilde{\Gamma}_i\supset\Gamma_i$ and $\widetilde{b}_i^v=\widetilde{b}_i^s+\widetilde{b}_i^p\epsilon<b_i^v$. In terms of Algorithm 3, we aim to achieve the *critical payment* property, i.e., each winner w_i is paid the supremum of the *virtual bidding price*, defined as $b_i^v=b_i^s+b_i^p\epsilon$, that can still make her a winner. As will be proved in Theorem 4, these two properties help ensure the truthfulness of the proposed pSRC auction.

Algorithm 2 pSRC Auction Winner Determination

Input: ϵ, $\mathbf{b}$, $\mathbf{q}$, $\mathbf{Q}$, $\mathcal{N}$, $\mathcal{T}$;
Output: $\mathcal{S}$;
// Initialization
1 $\mathcal{S}\leftarrow\emptyset$, $\mathbf{Q}'\leftarrow\mathbf{Q}$;
// Calculate the winner set $\mathcal{S}$
2 **while** $\sum_{j:\tau_j\in\mathcal{T}}Q_j'\neq 0$ **do**
  // Find the worker with the minimum bidding price effectiveness
3 $l=\arg\min_{i:w_i\in\mathcal{N}}\frac{b_i^s+b_i^p\epsilon}{\sum_{j:\tau_j\in\Gamma_i}\min\{Q_j',q_{i,j}\}}$;
4 $\mathcal{S}\leftarrow\mathcal{S}\cup\{w_l\}$;
5 $\mathcal{N}\leftarrow\mathcal{N}\setminus\{w_l\}$;
  // Update the $\mathbf{Q}'$ vector
6 **foreach** j *s.t.* $\tau_j\in\mathcal{T}$ **do**
7 $Q_j'\leftarrow Q_j'-\min\{Q_j',q_{l,j}\}$;
8 **return** $\mathcal{S}$;

The inputs of the winner determination algorithm given in Algorithm 2 include the privacy budget ϵ, bid profile $\mathbf{b}$, $\mathbf{q}$ matrix, $\mathbf{Q}$ vector, worker set $\mathcal{N}$, and task set $\mathcal{T}$. Firstly, it initializes the winner 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). In each iteration, Algorithm 2 finds the worker w_l 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 (19). Next, w_l 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.

Algorithm 3 pSRC Auction Pricing

Input: ϵ, $\mathbf{b}$, $\mathbf{q}$, $\mathbf{Q}$, $\mathcal{N}$, $\mathcal{T}$, $\mathcal{S}$;
Output: $\mathbf{p}$;
// Initialization
1 $\mathbf{p} \leftarrow (0, \cdots, 0)$;
2 **foreach** i *s.t.* $w_i \in \mathcal{S}$ **do**
3 run Algorithm 2 on $\mathcal{N} \setminus \{w_i\}$ until $\sum_{j:\tau_j \in \Gamma_i} Q'_j = 0$;
4 $\mathcal{S}' \leftarrow$ the winner set when step 3 stops;
// Calculate payment
5 **foreach** k *s.t.* $w_k \in \mathcal{S}'$ **do**
6 $\mathbf{Q}' \leftarrow$ tasks' $\mathbf{Q}'$ vector when w_k is selected;
7 $p_i \leftarrow \max\left\{p_i, (b_k^s + b_k^p\epsilon) \cdot \frac{\sum_{j:\tau_j \in \Gamma_i} \min\{Q'_j, q_{i,j}\}}{\sum_{j:\tau_j \in \Gamma_k} \min\{Q'_j, q_{k,j}\}}\right\}$;
8 **return** $\mathbf{p}$;

Besides the same inputs of Algorithm 2, the pricing algorithm in Algorithm 3 also uses the winner set $\mathcal{S}$ calculated by Algorithm 2. 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 winner. For each $w_i \in \mathcal{S}$, Algorithm 2 is executed on the worker set with 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

$$\frac{b_{i,k}^v}{\sum_{j:\tau_j \in \Gamma_i} \min\{Q'_j, q_{i,j}\}} = \frac{b_k^s + b_k^p\epsilon}{\sum_{j:\tau_j \in \Gamma_k} \min\{Q'_j, q_{k,j}\}},$$

which is equivalent to

$$b_{i,k}^v = (b_k^s + b_k^p\epsilon) \cdot \frac{\sum_{j:\tau_j \in \Gamma_i} \min\{Q'_j, q_{i,j}\}}{\sum_{j:\tau_j \in \Gamma_k} \min\{Q'_j, q_{k,j}\}}.$$

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

3) Analysis: Firstly, we analyze the truthfulness of the proposed pSRC auction in Theorem 4.

Theorem 4: The proposed pSRC auction is truthful.

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

- *Monotonicity:* Consider worker w_i's bid $\widetilde{b}_i = (\widetilde{\Gamma}_i, \widetilde{b}_i^s, \widetilde{b}_i^p)$ with $\widetilde{\Gamma}_i \supset \Gamma_i$ and $\widetilde{b}_i^v = \widetilde{b}_i^s + \widetilde{b}_i^p\epsilon < b_i^v$. Algorithm 2 selects winners in an increasing order of the bidding price effectiveness. Hence, $\widetilde{b}_i$ will also make worker w_i a winnner, 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 [15] and [42], the monotonicity and critical payment property make the pSRC auction truthful in terms of the bidding bundle and the virtual bidding price. That is worker w_i maximizes her utility by bidding Γ_i^* and (b_i^s, b_i^p) such that $b_i^s + b_i^p\epsilon = c_i^s + c_i^p\epsilon$. For a fixed ϵ, the worker still has incentive to bid $(b_i^s, b_i^p) \neq (c_i^s, c_i^p)$. However, since the exact value of ϵ is not revealed to workers in the bidding process, the only strategy that maximizes her utility under all possible values of ϵ is to bid $b_i^s = c_i^s$ and $b_i^p = c_i^p$. Therefore, the pSRC auction is truthful. □

The proposed pSRC auction ensures that truthful bidding is a dominant strategy for every worker under any possible value of ϵ. As stated in the proof of Theorem 4, it is crucial to keep the exact value of the privacy budget ϵ confidential to workers in the bidding process to ensure the truthfulness of a worker's bidding prices for the costs of sensing and unit privacy leakage, i.e., to achieve $b_i^s = c_i^s$ and $b_i^p = c_i^p$ for every worker w_i. The reason that the platform firstly announces to workers an upper bound of ϵ is to avoid their concerns of the possibility for excessively large privacy leakage. Next, we analyze the individual rationality of the pSRC auction.

Theorem 5: The pSRC auction is individual rational.

Proof: By Definition 5, losers of the auction receive zero utilities. From Theorem 4, every winner w_i bids to the platform the true value (c_i^s, c_i^p) 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 $p_i \geq c_i^s + c_i^p\epsilon$, 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 ϵ 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 6: The computational complexity of the proposed pSRC auction is $O(N^3 + N^2\ K)$.

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 $\mathbf{Q}'$ vector (line 6-7). Therefore, the computational complexity of Algorithm 2 is $O(N^2 + NK)$.

Furthermore, the computational complexity of Algorithm 3 is $O(N^3 + N^2\ K)$, 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 $O(N^3 + N^2\ K)$. □

Before analyzing the approximation ratio of the platform's total payment generated by the pSRC auction to the optimal total payment, we introduce Lemma 3 and 4 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:

$$\min \quad \sum_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p\epsilon) y_i \tag{25}$$

$$\text{s.t.} \quad \sum_{i:w_i \in \mathcal{N}, \tau_j \in \Gamma_i} q_{i,j} y_i \geq Q_j, \quad \forall \tau_j \in \mathcal{T} \tag{26}$$

$$y_i \in \{0, 1\}, \quad \forall w_i \in \mathcal{N} \tag{27}$$

The pSRC-SCM problem has the same set of inputs, constraints (including the inherent truthfulness and individual rationality 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., $\sum_{i:w_i \in \mathcal{S}} (c_i^s + c_i^p \epsilon)$, which is the sum of all winners' costs.

Lemma 3: The optimal social cost of the pSRC-SCM problem, denoted as 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 $P_{\text{OPT}} = \sum_{i:w_i \in \mathcal{N}} p_i^* y_i^*$.

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

$$C_{\text{OPT}} \leq \sum_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p \epsilon) y_i^* \leq \sum_{i:w_i \in \mathcal{N}} p_i^* y_i^* = P_{\text{OPT}},$$

which means that C_{OPT} is a lower bound of P_{OPT}. □

Then, we introduce Lemma 4 which is borrowed from Theorem 5 in [5] with some minor adaptations. Similar to [15], we introduce the following notations including $\eta = \max_{i,j:w_i \in \mathcal{N}, \tau_j \in \mathcal{T}} (c_i^s + c_i^p \epsilon) q_{i,j} |\Gamma_i|$ and $m = \frac{1}{\Delta q} \sum_{j:\tau_j \in \mathcal{T}} Q_j$ where Δq is the unit measure of elements in $\mathbf{q}$ and $\mathbf{Q}$.

Lemma 4: The social cost generated by Algorithm 2 satisfies $2\gamma H_m$-approximation to the optimal social cost, i.e.,

$$\sum_{i:w_i \in \mathcal{S}} (c_i^s + c_i^p \epsilon) \leq 2\eta H_m C_{\text{OPT}},$$

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

The proof to Lemma 4, which can be found in [15] is omitted in this paper. We define $\nu = \max_{i,k:w_i,w_k \in \mathcal{N}} \frac{c_i^s + c_i^p \epsilon}{c_k^s + c_k^p \epsilon}$, $\rho = \frac{1}{\Delta q} \max_{i,j:w_i \in \mathcal{N}, \tau_j \in \mathcal{T}} q_{i,j} |\Gamma_i|$, and introduce the following Theorem 7 regarding the approximation ratio of the proposed pSRC auction in terms of the platform's total payment.

Theorem 7: The platform's total payment generated by the proposed pSRC auction satisfies $2\rho\nu\eta H_m$-approximation to the optimal total payment, i.e.,

$$\sum_{i:w_i \in \mathcal{S}} p_i \leq 2\rho\nu\eta H_m P_{\text{OPT}}.$$

Proof: Based on Algorithm 3, for every winner w_i there exists some worker w_{k_i} such that

$$p_i = (c_{k_i}^s + c_{k_i}^p \epsilon) \cdot \frac{\sum_{j:\tau_j \in \Gamma_i} \min\{Q'_j, q_{i,j}\}}{\sum_{j:\tau_j \in \Gamma_{k_i}} \min\{Q'_j, q_{k_i,j}\}},$$

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

$$\begin{aligned}
\sum_{i:w_i \in \mathcal{S}} p_i &= \sum_{i:w_i \in \mathcal{S}} (c_{k_i}^s + c_{k_i}^p \epsilon) \cdot \frac{\sum_{j:\tau_j \in \Gamma_i} \min\{Q'_j, q_{i,j}\}}{\sum_{j:\tau_j \in \Gamma_{k_i}} \min\{Q'_j, q_{k_i,j}\}} \\
&\leq \max_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p \epsilon) \cdot \left(\frac{1}{\Delta q} \sum_{i:w_i \in \mathcal{S}} \sum_{j:\tau_j \in \Gamma_i} q_{i,j} \right) \\
&\leq |\mathcal{S}| \max_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p \epsilon) \cdot \left(\frac{1}{\Delta q} \max_{i,j:w_i \in \mathcal{N}, \tau_j \in \mathcal{T}} q_{i,j} |\Gamma_i| \right) \\
&= \rho |\mathcal{S}| \max_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p \epsilon). \qquad (28)
\end{aligned}$$

Furthermore, the social cost satisfies that

$$\sum_{i:w_i \in \mathcal{S}} (c_i^s + c_i^p \epsilon) \geq |\mathcal{S}| \min_{i:w_i \in \mathcal{N}} (c_i^s + c_i^p \epsilon). \qquad (29)$$

From Inequality (28) and (29), and Lemma 3 and 4, we have that

$$\begin{aligned}
\sum_{i:w_i \in \mathcal{S}} p_i &\leq \rho \left(\max_{i,k:w_i,w_k \in \mathcal{N}} \frac{c_i^s + c_i^p \epsilon}{c_k^s + c_k^p \epsilon} \right) \sum_{i:w_i \in \mathcal{S}} (c_i^s + c_i^p \epsilon) \\
&= \rho\nu \sum_{i:w_i \in \mathcal{S}} (c_i^s + c_i^p \epsilon) \leq 2\rho\nu\eta H_m C_{\text{OPT}} \\
&\leq 2\rho\nu\eta H_m P_{\text{OPT}}.
\end{aligned}$$

Therefore, the proposed pSRC auction satisfies $2\rho\nu\eta H_m$-approximation to the optimal total payment. □

Note that there is a $\max_{i \in \mathcal{N}} |\Gamma_i|$ factor in ρ and η, which could be large theoretically, and in worst case equals to the number of tasks K. However, practically, as a worker w_i typically has a limited capability and interest in terms of the number of tasks she can and wants to execute, $\max_{i \in \mathcal{N}} |\Gamma_i|$ will be far less than K, which prevents ρ and η from growing excessively large, in practice, as K increases. Furthermore, as $m = O(K)$ and $H_m = O(\log m)$, we have that $H_m = O(\log K)$. Therefore, although H_m is not a constant, it is still much smaller than K in order sense. Thus far, the $2\rho\nu\eta H_m$ approximation ratio proved in Theorem 7 is the best one we have found, and we leave the proof of the tightness of this ratio, or the derivation of a better one in our future work.

C. Data Perturbation Mechanism

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 (8). That is, it has to know α and θ, which is usually feasible for the adversary platform. For similar sensing tasks, α is typically a common and standard design choice across different platforms, and workers' reliability levels for similar tasks tend to be similar as well. Therefore, θ can also be effectively estimated or inferred by the adversary platform using the methods mentioned in Section III-C, such as utilizing workers' sensory data about similar tasks collected during its past interactions with them as in [39] and [40], using some of workers' characteristics (e.g., reputation and experience for similar tasks) as in [41], and many others. To tackle such inference attack, we propose a novel data perturbation mechanism in Algorithm 4 by tailoring the Laplace mechanism in [26] and [43] to our problem setting.

Apart from the the vector of the aggregated results $(x_1, \cdots, x_N)$ output by the data aggregation mechanism, the task set $\mathcal{T}$, the same α and β vector as in Algorithm 1, 2, and 3, Algorithm 4 also takes as input the vector $\widetilde{\mathbf{x}}$, where each element $\widetilde{x}_j$ corresponds to one categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$ with

$$\widetilde{x}_j = \sum_{i:w_i \in \mathcal{S}, \tau_j \in \Gamma_i} (2\theta_{i,j} - 1) x_{i,j}. \qquad (30)$$

Algorithm 4 Data Perturbation Mechanism

Input: $(x_1, \cdots, x_N)$, α, β, $\mathcal{T}$, $\widetilde{\mathbf{x}}$, δ;
Output: $(\widehat{x}_1, \cdots, \widehat{x}_N)$;
1 **foreach** j *s.t.* $\tau_j \in \mathcal{T}$ **do**
2 **if** $\tau_j \in \mathcal{T}_{con}$ **then**
3 randomly sample a noise n_j from Lap$\Big(0, -\frac{\alpha_j}{\ln \beta_j}\Big)$;
4 $\widehat{x}_j \leftarrow x_j + n_j$;
5 **else**
6 randomly sample a noise n_j from Lap$\Big(0, \frac{1}{\delta_j}\Big)$;
7 $\widehat{x}_j \leftarrow \text{sign}(\widetilde{x}_j + n_j)$;
8 **return** $(\widehat{x}_1, \cdots, \widehat{x}_N)$;

Clearly, for each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, $\widetilde{x}_j$ is its intermediate aggregated result before we convert it to the binary label x_j. Although not explicitly described, Algorithm 1 keeps track of these intermediate results $\widetilde{x}_j$'s so that they can be utilized by Algorithm 4. Additionally, the last input parameter to Algorithm 4 is the vector δ, where each element $\delta_j \in (0, 1)$ is a platform-chosen parameter corresponding to the privacy guarantee of a categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$. For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, 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$\big(0, -\frac{\alpha_j}{\ln \beta_j}\big)$ (line 3), and adds it to the aggregated result x_j (line 4). For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, the algorithm randomly samples a noise from the Laplacian distribution with mean 0 and scaling $\frac{1}{\delta_j}$, denoted as Lap$\big(0, \frac{1}{\delta_j}\big)$ (line 6), and the perturbed result $\widehat{x}_j$ of this task is calculated as sign$(\widetilde{x}_j + n_j)$ (line 7). Although adding Laplacian noise as in [26] and [43] 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.

2) Analysis: We firstly analyze Algorithm 4's accuracy guarantee for continuous tasks.

Theorem 8: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, the data perturbation mechanism given in Algorithm 4 satisfies

$$\Pr\big[\big|\widehat{X}_j - X_j\big| \geq \alpha_j\big] = \beta_j. \tag{31}$$

Proof: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, we use N_j to denote the random variable representing the random noise sampled from the Laplacian distribution Lap$\big(0, -\frac{\alpha_j}{\ln \beta_j}\big)$, i.e., $N_j \sim \text{Lap}\big(0, -\frac{\alpha_j}{\ln \beta_j}\big)$. Thus,

$$\begin{aligned}\Pr\big[\big|\widehat{X}_j - X_j\big| \geq \alpha_j\big] &= \Pr\big[\big|N_j\big| \geq \alpha_j\big] = 2\Pr\big[N_j \geq \alpha_j\big] \\ &= 2\int_{\alpha_j}^{+\infty} -\frac{\ln \beta_j}{2\alpha_j} \exp\left(\frac{z \ln \beta_j}{\alpha_j}\right) dz = \beta_j,\end{aligned}$$

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

Theorem 8 states that (α_j, β_j)-accuracy is guaranteed for the perturbed result compared to the original one before perturbation for every continuous task $\tau_j \in \mathcal{T}_{\text{con}}$. 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 9.

Theorem 9: For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, the data perturbation mechanism given in Algorithm 4 satisfies

$$\Pr\big[\big|\widehat{X}_j - x_j^*\big| \geq 2\alpha_j\big] \leq 1 - (1 - \beta_j)^2. \tag{32}$$

Proof: As discussed in Section IV-A and IV-B, the aggregated result for every continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ satisfies that $\Pr\big[\big|X_j - x_j^*\big| \geq \alpha_j\big] \leq \beta_j$. From Theorem 8 and the fact that $X_j - x_j^*$ and $\widehat{X}_j - X_j = N_j$ are two independent random variables, we have

$$\begin{aligned}\Pr\big[\big|\widehat{X}_j - x_j^*\big| > 2\alpha_j\big] &\leq \Pr\big[\big|\widehat{X}_j - X_j\big| + \big|X_j - x_j^*\big| > 2\alpha_j\big] \\ &\leq 1 - (1 - \beta_j)^2,\end{aligned}$$

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

Therefore, Theorem 9 gives us that $(2\alpha_j, 1 - (1 - \beta_j)^2)$-accuracy is satisfied for the perturbed result of every continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ compared to its ground truth. Next, we analyze Algorithm 4's accuracy guarantee for categorical tasks.

Theorem 10: For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, the data perturbation mechanism given in Algorithm 4 satisfies

$$\Pr\big[\widehat{X}_j \neq x_j^*\big] \leq \frac{\gamma_j + 1}{2}. \tag{33}$$

Proof: For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, we have that

$$\begin{aligned}\Pr\big[\widehat{X}_j \neq x_j^*\big] = {} & \Pr\big[\widetilde{X}_j + N_j \geq 0 | x_j^* = -1\big]\Pr\big[x_j^* = -1\big] \\ & + \Pr\big[\widetilde{X}_j + N_j < 0 | x_j^* = +1\big]\Pr\big[x_j^* = +1\big],\end{aligned}$$

where $\widehat{X}_j$ denotes the random variable corresponding to $\widetilde{x}_j$, and N_j denotes the random variable that represents the random noise sampled from the Laplacian distribution Lap$\big(0, \frac{1}{\gamma_j}\big)$. Then, we have that

$$\begin{aligned}&\Pr\big[\widetilde{X}_j + N_j \geq 0 | x_j^* = -1\big] \\ &\qquad \leq 1 - \Pr\big[\widetilde{X}_j < 0 | x_j^* = -1\big]\Pr\big[N_j < 0\big] \\ &\qquad < 1 - \frac{1 - \gamma_j}{2} = \frac{1 + \gamma_j}{2},\end{aligned}$$

where the last inequality is because of $\Pr\big[\widetilde{X}_j < 0 | x_j^* = -1\big] > 1 - \gamma_j$ which is an intermediate result in the proof of [14, Th. 1]. Similarly, we have that $\Pr\big[\widetilde{X}_j + N_j < 0 | x_j^* = +1\big] < \frac{1+\gamma_j}{2}$. Therefore, we have that

$$\Pr\big[\widehat{X}_j \neq x_j^*\big] \leq \frac{\gamma_j + 1}{2}, \tag{34}$$

which exactly proves this Theorem. □

By Theorem 10, we have that the final perturbed result of each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$ satisfies γ_j-accuracy compared to its ground truth with $\gamma_j \in (0, 1)$. Next, in Theorem 11, we analyze the privacy guarantee of the data perturbation mechanism.

Theorem 11: The data perturbation mechanism given in Algorithm 4 satisfies ϵ-differential privacy, where the privacy budget $\epsilon = \max\Big\{\max_{j:\tau_j \in \mathcal{T}_{\text{con}}}\Big(-\frac{\ln \beta_j}{\alpha_j}\Big),\ \max_{j:\tau_j \in \mathcal{T}_{\text{cat}}} 2\delta_j\Big\}$.

Proof: Similar to the proof of Theorem 8 and 10, we use N_j to denote the random variable corresponding to the random noise n_j sampled by Algorithm 4 for each task τ_j. 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 $x_j^{(i)}$ and $\widehat{x}_j^{(i)}$ to denote the aggregated result for task τ_j before and after perturbation when one worker w_i's

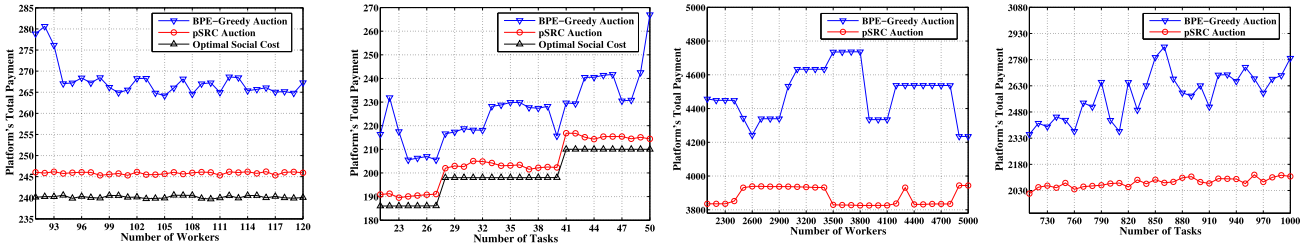


Fig. 2. Platform's total payment for (a) setting I, (b) setting II, (c) setting III, and (d) setting IV.

data $x_{i,j}$ changes. For each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$, we have $\left|x_j - x_j^{(i)}\right| \leq 1$, and

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

For each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$, we use $\widetilde{x}_j^{(i)}$ to denote the value of $\widetilde{x}_j$ when one worker w_i's data $x_{i,j}$ changes, and clearly $\left|x_j - x_j^{(i)}\right| \leq 2$. Thus, we have that

$$\begin{aligned}
\Pr\left[\widetilde{X}_j + N_j \in \mathcal{O}\right] &= \Pr\left[N_j \in \mathcal{O} - \widetilde{X}_j\right] \\
&= \int_{z \in \mathcal{O} - \widetilde{X}_j} \frac{\delta_j}{2} \exp\left(-\delta_j |z|\right) dz \\
&\leq \exp\left(2\delta_j\right) \int_{z \in \mathcal{O} - \widetilde{X}_j^{(i)}} \frac{\delta_j}{2} \exp\left(-\delta_j |z|\right) dz \\
&= \exp\left(2\delta_j\right) \Pr\left[\widetilde{X}_j^{(i)} + N_j \in \mathcal{O}\right].
\end{aligned}$$

As $\mathcal{O}$ could be any subset of $\mathbb{R}$, we let $\mathcal{O} = [0, +\infty)$, and get $\Pr[\widetilde{X}_j + N_j \geq 0] \leq \exp(2\delta_j) \Pr[\widetilde{X}_j^{(i)} + N_j \geq 0]$. Thus, we have that

$$\frac{\Pr\left[\widehat{X}_j = +1\right]}{\Pr\left[\widehat{X}_j^{(i)} = +1\right]} = \frac{\Pr\left[\widetilde{X}_j + N_j \geq 0\right]}{\Pr\left[\widetilde{X}_j^{(i)} + N_j \geq 0\right]} \leq \exp(2\delta_j).$$

Similarly, by letting $\mathcal{O} = (-\infty, 0)$, we have that

$$\frac{\Pr\left[\widehat{X}_j = -1\right]}{\Pr\left[\widehat{X}_j^{(i)} = -1\right]} = \frac{\Pr\left[\widetilde{X}_j + N_j < 0\right]}{\Pr\left[\widetilde{X}_j^{(i)} + N_j < 0\right]} \leq \exp(2\delta_j).$$

Note that the previous analysis focuses on a specific task τ_j. The overall privacy budget considering all tasks in $\mathcal{T}$ is thus $\epsilon = \max\left\{\max_{j:\tau_j \in \mathcal{T}_{\text{con}}}\left(-\frac{\ln \beta_j}{\alpha_j}\right), \max_{j:\tau_j \in \mathcal{T}_{\text{cat}}} 2\delta_j\right\}$. □

D. Summary of Design Details

Thus far, we have finished the description of the design details of INCEPTION. Its incentive mechanism (Section IV-B) 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 7), and satisfies computational efficiency (Theorem 6), truthfulness (Theorem 4), and individual rationality (Theorem 5). Incorporating workers' reliability levels, the data aggregation mechanism (Section IV-A) provides aggregated results with high accuracy (Corollary 1 and 2), and the data perturbation mechanism (Section IV-C) adds carefully controlled noises to the aggregated results to achieve differential privacy (Theorem 11), and small degradation of aggregation accuracy (Theorem 8, 9, and 10).

Overall, INCEPTION guarantees $\max\left\{\max_{j:\tau_j \in \mathcal{T}_{\text{con}}}\left(-\frac{\ln \beta_j}{\alpha_j}\right), \max_{j:\tau_j \in \mathcal{T}_{\text{cat}}} 2\delta_j\right\}$-differential privacy, $(2\alpha_j, 1 - (1-\beta_j)^2)$-accuracy for each continuous task $\tau_j \in \mathcal{T}_{\text{con}}$ (Theorem 9), and $\frac{\gamma_j+1}{2}$-accuracy for each categorical task $\tau_j \in \mathcal{T}_{\text{cat}}$ (Theorem 10). The platform could carefully select $\alpha_j, \beta_j, \gamma_j, \delta_j \in (0,1)$ for each task τ_j to ensure satisfactory guarantee for aggregation accuracy and workers' privacy.

V. Performance Evaluation

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

A. 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 [44], [45] 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 3. 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 an increasing order of 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 baseline aggregation methods, namely the mean and median aggregation. For each continuous task, the mean and median aggregation method simply utilizes, respectively, the mean and median of workers' data as its aggregated result. For each categorical task, the median aggregation method also uses the median of workers' data as the task's aggregated result, but the mean aggregation method firstly calculates the mean of workers' data about this task, and then takes the sign of the mean as the aggregated result.

TABLE I

SIMULATION SETTINGS (CONTINUOUS TASKS ONLY)

Setting	α_j, β_j	c_i^s, c_i^p	$\mu_{i,j}$, x_j^*	$\sigma_{i,j}$	$\lvert\Gamma_i^*\rvert$	N	K
I	$(0, 0.1]$	$[1, 2]$	$[0, 1]$	$[1, 2]$	$[15, 20]$	$[91, 120]$	40
II	$(0, 0.1]$	$[1, 2]$	$[0, 1]$	$[1, 2]$	$[15, 20]$	100	$[21, 50]$
III	$(0, 0.1]$	$[1, 2]$	$[0, 1]$	$[1, 2]$	$[25, 35]$	$[2100, 5000]$	500
IV	$(0, 0.1]$	$[1, 2]$	$[0, 1]$	$[1, 2]$	$[25, 35]$	1000	$[710, 1000]$

TABLE II

SIMULATION SETTINGS (CATEGORICAL TASKS ONLY)

Setting	γ_j, δ_j	c_i^s, c_i^p	x_j^*	$\theta_{i,j}$	$\lvert\Gamma_i^*\rvert$	N	K
V	$(0, 0.1]$	$[1, 2]$	$\{-1, +1\}$	$(0, 1)$	$[15, 20]$	$[91, 120]$	40
VI	$(0, 0.1]$	$[1, 2]$	$\{-1, +1\}$	$(0, 1)$	$[15, 20]$	100	$[21, 50]$
VII	$(0, 0.1]$	$[1, 2]$	$\{-1, +1\}$	$(0, 1)$	$[25, 35]$	$[2100, 5000]$	500
VIII	$(0, 0.1]$	$[1, 2]$	$\{-1, +1\}$	$(0, 1)$	$[25, 35]$	1000	$[710, 1000]$

B. Simulation Settings

For simplicity of presenting our simulation results, we consider setting I-IV in Table I where the platform hosts only continuous tasks, and Setting V-VIII where the platform hosts only categorical tasks. Note that, clearly, our INCEPTION framework is applicable in the scenario where both continuous and categorical tasks are hosted by the platform.

For each continuous task τ_j, we generate worker w_i's data about this task, i.e., $x_{i,j}$, from a normal distribution with mean $\mu_{i,j}$ and standard deviation $\sigma_{i,j}$, truncated within the range $[0, 1]$. The value of $\theta_{i,j}$ for each continuous task τ_j is calculated by platform as

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

where $c_{i,j} = \left(\Phi\left(\frac{1-\mu_{i,j}}{\sigma_{i,j}}\right) - \Phi\left(-\frac{\mu_{i,j}}{\sigma_{i,j}}\right)\right)^{-1}$, $b_{i,j} = \mu_{i,j} - x_j^*$, $a_{i,j} = x_j^* + b_{i,j}$, and $\Phi(\cdot)$ 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 for the scenarios with only continuous tasks are given in Table I.

In setting I and II, α_j, β_j, c_i^s, c_i^p, x_j^*, $\mu_{i,j}$, $\sigma_{i,j}$, and $\lvert\Gamma_i^*\rvert$ are generated uniformly at random from the intervals given in Table I. The bundle Γ_i^* contains $\lvert\Gamma_i^*\rvert$ tasks randomly chosen from $\mathcal{T}$. In setting I, we fix the number of tasks as 40 and vary the number of workers from 91 to 120. In contrast, we fix the number of workers as 100 and vary the number of tasks from 21 to 50 in setting II. In setting III and IV, α_j, β_j, c_i^s, c_i^p, x_j^*, $\mu_{i,j}$, $\sigma_{i,j}$, and $\lvert\Gamma_i^*\rvert$ are generated in the same way as in setting I and II from the intervals given in Table I. Different from the previous two settings, setting III and IV take instances with larger sizes, given in Table I, as inputs. Next, we give our parameter settings for the scenarios with only categorical tasks in Table II.

In setting V and VI, γ_j, δ_j, c_i^s, c_i^p, x_j^*, $\theta_{i,j}$, and $\lvert\Gamma_i^*\rvert$ are generated uniformly at random from the intervals given in Table II. The bundle Γ_i^* contains $\lvert\Gamma_i^*\rvert$ tasks randomly chosen from $\mathcal{T}$. In setting V, we fix the number of tasks at 40 and vary the number of workers from 91 to 120. In contrast, we fix the number of workers at 100 and vary the number of tasks from 21 to 50 in setting VI. In setting VII and VIII, the parameters γ_j, δ_j, c_i^s, c_i^p, x_j^*, $\theta_{i,j}$, and $\lvert\Gamma_i^*\rvert$ are generated in the same way as in setting V and VI from the intervals given in Table II. Different from the previous two settings, setting VII and VIII take instances with larger sizes as inputs, which are given in Table II. The optimal solutions to the pSRC-SCM problem are calculated using the GUROBI optimization solver [46].

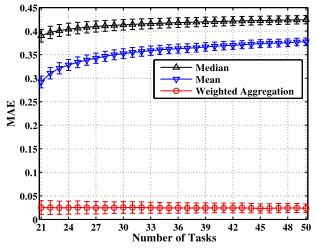


Fig. 3. MAE of data aggregation (continuous tasks only).

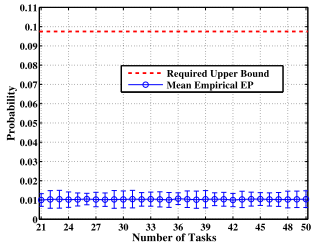


Fig. 4. EP after perturbation (continuous tasks only).

TABLE III

EXECUTION TIME (SECONDS) FOR SETTING I AND II

N	91	95	99	103	107	111	115	119
VCG	20.23	79.11	227.5	257.7	308.7	836.4	1199	1537
pSRC	0.008	0.009	0.007	0.008	0.008	0.006	0.007	0.006
K	21	25	29	33	37	41	45	49
VCG	0.300	6.676	13.09	30.60	1063	1160	1330	1677
pSRC	0.003	0.005	0.003	0.007	0.009	0.009	0.003	0.003

TABLE IV

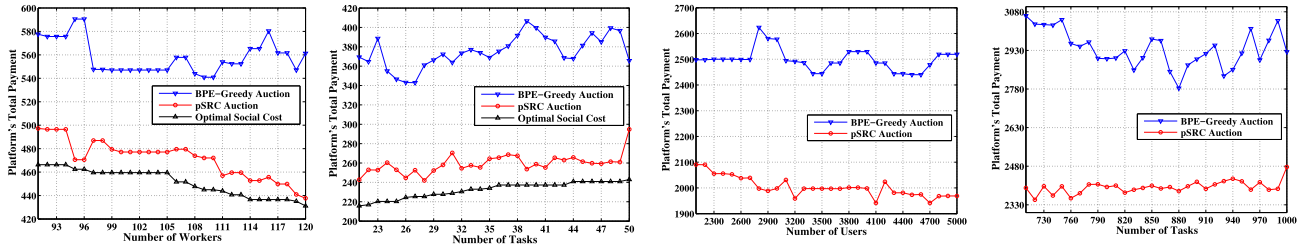
EXECUTION TIME (SECONDS) FOR SETTING V AND VI

N	91	95	99	103	107	111	115	119
VCG	7.397	25.31	115.5	225.6	312.4	517.4	1059	1105
pSRC	0.016	0.018	0.018	0.019	0.020	0.019	0.021	0.024
K	21	25	29	33	37	41	45	49
VCG	5.400	16.33	33.90	500.4	735.9	1050	1100	1507
pSRC	0.016	0.017	0.019	0.019	0.020	0.021	0.023	0.024

C. Simulation Results

Figure 2-4 demonstrate our simulation results on Setting I-IV with only continuous tasks. Figure 2(a) and (b) 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 III, 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



(a) (b) (c) (d)

Fig. 5. Platform's total payment for (a) setting V, (b) setting VI, (c) setting VII, and (d) setting VIII.

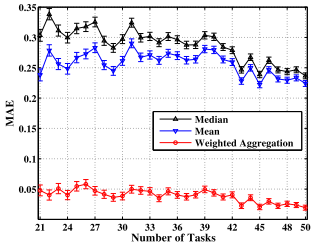


Fig. 6. MAE of data aggregation (categorical tasks only).

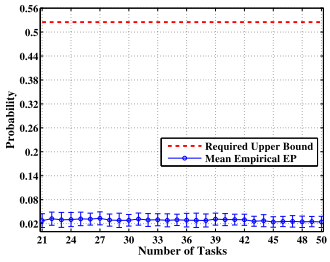


Fig. 7. EP after perturbation (categorical tasks only).

pSRC auction keeps in the order of microsecond, which is much less that of the VCG auction.

In Figure 2(c) and (d), 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 less than that of the BPE-Greedy auction.

We evaluate the accuracy guarantee of INCEPTION in setting II with a minor change of the parameter β_j, i.e., β_j is fixed as 0.05 for every task τ_j to simplify presentation. We compare the mean absolute error (MAE) for all tasks, defined as MAE $= \frac{1}{K}\sum_{j:\tau_j\in\mathcal{T}}|x_j - x_j^*|$, 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 3 that the MAE of our weighted aggregation is far less than those of the mean and median aggregation. Then, we show simulation results regarding $\Pr[|\widehat{X}_j - x_j^*| \geq \alpha_j]$, referred to as the error probability (EP) of the perturbed results for task τ_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 4, we observe that the empirical EPs are far less than the required upper bound (i.e., $1-(1-\beta_j)^2 = 1-(1-0.05)^2 = 0.0975$).

Next, we show our simulation results for setting V-VIII with only categorical tasks in Figure 5-7, which share similar trends as Figure 2-4. The simulation setting for Figure 6 and 7 is the same as setting IV except that γ_j and δ_j for each task τ_j are fixed as 0.1. In Figure 7, EP is defined as $\Pr[\widehat{X}_j \neq x_j^*]$, whose empirical value is calculated in the same way as that for a continuous task in Figure 4.

Furthermore, we show in Table IV the comparison between the execution time of the VCG and the BPE-Greedy auction for setting V and VI. Clearly, similar to Table III, Table IV also shows that execution time of the pSRC auction is much less that of the VCG auction.

VI. Conclusion

In this paper, 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.

Acknowledgment

The authors thank Professor R. Srikant for his valuable comments.

References

[1] R. Gao *et al.*, "Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing," in *Proc. MobiCom*, 2014, pp. 249–260.
[2] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "GreenGPS: A participatory sensing fuel-efficient maps application," in *Proc. MobiSys*, 2010, pp. 151–164.
[3] S. Hu, L. Su, H. Liu, H. Wang, and T. F. Abdelzaher, "SmartRoad: Smartphone-based crowd sensing for traffic regulator detection and identification," *ACM Trans. Sensor Netw.*, vol. 11, no. 4, 2015, Art. no. 55.
[4] J. Eriksson *et al.*, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Proc. MobiSys*, 2008, pp. 29–39.
[5] Y. Chen, B. Li, and Q. Zhang, "Incentivizing crowdsourcing systems with network effects," in *Proc. INFOCOM*, Apr. 2016, pp. 1–9.
[6] S. He, D.-H. Shin, J. Zhang, and J. Chen, "Toward optimal allocation of location dependent tasks in crowdsensing," in *Proc. INFOCOM*, Apr./May 2014, pp. 745–753.
[7] T. Luo, S. S. Kanhere, H.-P. Tan, F. Wu, and H. Wu, "Crowdsourcing with tullock contests: A new perspective," in *Proc. INFOCOM*, Apr./May 2015, pp. 2515–2523.
[8] L. Duan *et al.*, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proc. INFOCOM*, Mar. 2012, pp. 1701–1709.
[9] D. Peng, F. Wu, and G. Chen, "Pay as how well you do: A quality based incentive mechanism for crowdsensing," in *Proc. MobiHoc*, 2015, pp. 177–186.

[10] H. Xie, J. C. S. Lui, J. W. Jiang, and W. Chen, "Incentive mechanism and protocol design for crowdsourcing systems," in *Proc. Allerton*, Sep./Oct. 2014, pp. 140–147.
[11] K. Han, H. Huang, and J. Luo, "Posted pricing for robust crowdsensing," in *Proc. MobiHoc*, 2016, pp. 261–270.
[12] M. H. Cheung, R. Southwell, F. Hou, and J. Huang, "Distributed time-sensitive task selection in mobile crowdsensing," in *Proc. MobiHoc*, 2015, pp. 157–166.
[13] L. Gao, F. Hou, and J. Huang, "Providing long-term participation incentive in participatory sensing," in *Proc. INFOCOM*, Apr./May 2015, pp. 2803–2811.
[14] H. Jin, L. Su, and K. Nahrstedt, "CENTURION: Incentivizing multi-requester mobile crowd sensing," in *Proc. INFOCOM*, May 2017, pp. 1–9.
[15] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," in *Proc. MobiHoc*, 2015, pp. 167–176.
[16] Y. Wen *et al.*, "Quality-driven auction-based incentive mechanism for mobile crowd sensing," *IEEE Trans. Veh. Technol.*, vol. 64, no. 9, pp. 4203–4214, Sep. 2014.
[17] D. Zhao, X.-Y. Li, and H. Ma, "How to crowdsource tasks truthfully without sacrificing utility: Online incentive mechanisms with budget constraint," in *Proc. INFOCOM*, Apr./May 2014, pp. 1213–1221.
[18] Q. Zhang, Y. Wen, X. Tian, X. Gan, and X. Wang, "Incentivize crowd labeling under budget constraint," in *Proc. INFOCOM*, Apr./May 2015, pp. 2812–2820.
[19] X. Zhang, G. Xue, R. Yu, D. Yang, and J. Tang, "Truthful incentive mechanisms for crowdsourcing," in *Proc. INFOCOM*, Apr./May 2015, pp. 2830–2838.
[20] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proc. Mobicom*, 2012, pp. 173–184.
[21] D. Yang, G. Xue, X. Fang, and J. Tang, "Incentive mechanisms for crowdsensing: Crowdsourcing with smartphones," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1732–1744, Jun. 2016.
[22] X. Zhang *et al.*, "Free market of crowdsourcing: Incentive mechanism design for mobile sensing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3190–3200, Dec. 2014.
[23] Z. Feng, Y. Zhu, Q. Zhang, L. Ni, and A. Vasilakos, "TRAC: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing," in *Proc. INFOCOM*, Apr./May 2014, pp. 1231–1239.
[24] H. Zhang, B. Liu, H. Susanto, G. Xue, and T. Sun, "Incentive mechanism for proximity-based mobile crowd service systems," in *Proc. INFOCOM*, Apr. 2016, pp. 1–9.
[25] J. Wang, J. Tang, D. Yang, E. Wang, and G. Xue, "Quality-aware and fine-grained incentive mechanisms for mobile crowdsensing," in *Proc. ICDCS*, Jun. 2016, pp. 354–363.
[26] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. EC*, 2011, pp. 199–208.
[27] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. EC*, 2012, pp. 568–585.
[28] Q. Li and G. Cao, "Efficient privacy-preserving stream aggregation in mobile sensing with low aggregation error," in *Proc. PETS*, 2013, pp. 60–81.
[29] Q. Li and G. Cao, "Efficient and privacy-preserving data aggregation in mobile sensing," in *Proc. ICNP*, Oct./Nov. 2012, pp. 1–10.
[30] C. Miao *et al.*, "Cloud-enabled privacy-preserving truth discovery in crowd sensing systems," in *Proc. SenSys*, 2015, pp. 183–196.
[31] F. Qiu, F. Wu, and G. Chen, "Privacy and quality preserving multimedia data aggregation for participatory sensing systems," *IEEE Trans. Mobile Comput.*, vol. 14, no. 6, pp. 1287–1300, Jun. 2015.
[32] T. Jung *et al.*, "Privacy-preserving data aggregation without secure channel: Multivariate polynomial evaluation," in *Proc. INFOCOM*, Apr. 2013, pp. 2634–2642.
[33] W. Wang, L. Ying, and J. Zhang, "A game-theoretic approach to quality control for collecting privacy-preserving data," in *Proc. Allerton*, Sep./Oct. 2015, pp. 474–479.
[34] Q. Li and G. Cao, "Providing efficient privacy-aware incentives for mobile sensing," in *Proc. ICDCS*, Jun./Jul. 2014, pp. 208–217.
[35] Q. Li and G. Cao, "Providing privacy-aware incentives for mobile sensing," in *Proc. PerCom*, Mar. 2013, pp. 76–84.
[36] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *Proc. ICDCS*, Jun. 2016, pp. 344–353.
[37] J. Lin, D. Yang, M. Li, J. Xu, and G. Xue, "BidGuard: A framework for privacy-preserving crowdsensing incentive mechanisms," in *Proc. CNS*, Oct. 2016, pp. 145–153.
[38] D. Oleson *et al.*, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing," in *Proc. HCOMP*, 2011, pp. 43–48.
[39] Q. Li *et al.*, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *Proc. SIGMOD*, 2014, pp. 1187–1198.
[40] C. Meng *et al.*, "Truth discovery on crowd sensing of correlated entities," in *Proc. SenSys*, 2015, pp. 169–182.
[41] H. Li, B. Zhao, and A. Fuxman, "The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing," in *Proc. WWW*, 2014, pp. 165–176.
[42] L. Blumrosen and N. Nisan, "Combinatorial auctions," in *Algorithmic Game Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[43] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*. New York, NY, USA: Springer, 2011, pp. 338–340.
[44] E. H. Clarke, "Multipart pricing of public goods," *Public Choice*, vol. 11, no. 1, pp. 17–33, 1971.
[45] T. Groves, "Incentives in teams," *Econometrica*, vol. 41, no. 4, pp. 617–631, 1973.
[46] *Gurobi Solver*. [Online]. Available: http://www.gurobi.com/

Haiming Jin received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2012, and the Ph.D. degree from the University of Illinois at Urbana–Champaign (UIUC), Urbana, IL, USA, in 2017. He is currently a tenure-track Assistant Professor at the John Hopcroft Center for Computer Science and the Department of Electronic Engineering, Shanghai Jiao Tong University. Before this, he was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, UIUC. He is broadly interested in addressing unfolding research challenges in the general areas of urban computing, cyber-physical systems, crowd and social sensing systems, network economics and game theory, reinforcement learning, and mobile pervasive and ubiquitous computing.

Lu Su (M'15) received the M.S. degree in statistics and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, in 2012 and 2013, respectively. He was with the IBM T. J. Watson Research Center and the National Center for Supercomputing Applications. He is currently an Assistant Professor with the Department of Computer Science and Engineering, SUNY Buffalo. His research focuses on the general areas of mobile and crowd sensing systems, Internet of Things, and cyber-physical systems. He is a member of the ACM. He was a recipient of the NSF CAREER Award, the University at Buffalo Young Investigator Award, the ICCPS17 Best Paper Award, and the ICDCS17 Best Student Paper Award.

Houping Xiao received the B.S. degree in statistics from Beijing Normal University. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, SUNY Buffalo. His research interests are broadly in data mining and machine learning, including truth discovery, multi-source information trustworthiness analysis, privacy-preserving data mining, and distributed machine learning.

Klara Nahrstedt (S'93–M'95–SM'06–F'08) received the B.A. degree in mathematics and the M.Sc. degree in numerical analysis from the Humboldt University of Berlin, Berlin, Germany, in 1984 and 1985, respectively, and the Ph.D. degree from the Department of Computer and Information Science, University of Pennsylvania, in 1995. She is currently a Ralph and Catherine Fisher Professor with the Computer Science Department and the Director of the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign. She is an ACM Fellow and a member of the Leopoldina German National Academy of Sciences. She was a recipient of the IEEE Communication Society Leonard Abraham Award for research achievements, the Humboldt Award, and the IEEE Computer Society and ACM SIGMM Technical Achievement Awards.