Privacy Considerations in Participatory Data Collection via Spatial Stackelberg Incentive Mechanisms.

Jing Yang Koh  
*Department of Electrical & Computer Engineering,  
National University of Singapore,  
Singapore 117576.*

Gareth W. Peters  
*(corresponding author)*  
*Department of Actuarial Mathematics and Statistics  
Heriot-Watt,  
Edinburgh, EH14 4AS UK*

Ido Nevat  
*TUMCREATE Ltd  
1 CREATE Way, CREATE Tower,  
10-02 Singapore 138602.*

Derek Leong  
*Institute for Infocomm Research (IR)  
Agency for Science, Technology and Research (A*STAR),  
138632 Singapore.*

Editor: Francis Bach, David Blei and Bernhard Schölkopf

**Abstract**

Mobile crowd sensing is a widely used sensing paradigm allowing applications on mobile smart devices to routinely obtain spatially distributed data on a range of user attributes: location, temperature, video and audio. Such data then typically forms the input to application specific machine learning tasks to achieve objectives such as improving user experience, targeting geo-localised query based searches to user interests and commercial aspects of targeted geo-localised advertising.

We consider a scenario in which the sensing application purchases data from spatially distributed smartphone users. In many spatial monitoring applications, the crowdsourcer needs to incentivize users to contribute sensing data. This may help ensure collected data has good spatial coverage, which will enhance quality of service provided to the application user when used in machine learning tasks such as spatial regression.

Privacy considerations should be addressed in such crowd sensing applications, and an incentive offered to “privacy-concerned” users to contribute data. A novel Stackelberg incentive mechanism is developed that allows workers to specify their location whilst satisfying their location privacy requirements. The Stackelberg and Nash equilibria are explored and an algorithm to demonstrate the approach is developed for a real data application.

**Keywords:** Incentive mechanism design, Stackelberg game, location privacy, mobile crowd sensing, privacy.

1. Introduction

Mobile crowd sensing is becoming a widely used sensing paradigm as it allows applications on smart devices to routinely and efficiently obtain spatially distributed data on a diverse range of user attributes. Such data then typically forms the basis of input data in a variety
of machine learning tasks that could have objectives such as improving user experience of applications they run on their smartphone or mobile device, targeting query based searches to their particular interests and geographic location as well as the commercial aspect of targeted advertising.

Previous studies on crowd sourcing data in machine learning contexts, such as Shah et al. (2015) and Shah and Zhou (2016), focused on aspects of provision of labeled training data that is facilitated by crowdsourcing frameworks. In such works, the authors focused on interesting questions pertaining to incentive mechanisms and developed voting schemes to improve aspects of the data collection in such frameworks to address practical concerns such the quality of the crowd sourced labels. They argued that such settings were of relevance in two contexts. Settings in which the application users (typically termed workers) were performing crowdsourced data labeling when in fact they were not experts on the labeling task or alternatively when the application interface is not adequately designed to ensure that workers were able to convey their knowledge accurately on the labeling task.

In this work we consider different aspects of the crowdsourcing data provision framework that were not previously studied. However, we believe they are of direct practical relevance to numerous applications in this emerging area of crowdsourced data provision for machine learning algorithms. We consider the situation in which the sensing application may incentivise or “buy” sensor data from spatially distributed mobile smartphone users (workers) instead of deploying their own sensor networks. In such contexts, the geo-location of the sourced data is directly relevant to the machine learning task that the application will consider when providing the content to the users of the application. Furthermore, we consider an important aspect not previously addressed related to the ability to incentivise crowd sourced data in the presence of workers who are privacy aware and therefore require greater monetary incentives to forgo or relax their privacy concerns when supplying data to the crowd sourcing application.

The spatial aspect of such problems, where the geo-location of sourced data is of direct relevance, is previously under explored but is relevant to many unsupervised or semi-supervised machine learning application contexts. For instance, we explore a setting in which one may wish to perform a statistical task of say Gaussian process spatial field regression or estimation using crowdsourced data. Such tasks can then provide estimates of application specific information to a user at the location of the smartphone, based on a statistical models calibrated to previously sourced spatially distributed data of relevance. Numerous examples come to mind for such applications including localized climate information, traffic density information, local pollution levels etc.

Such mobile crowd sensing platforms Guo et al. (2015) have also become an emerging sensing paradigm in the age of the Internet of Things (IoT) that replaces fixed sensing infrastructure and removes its deployment and maintenance costs. The interface of such IoT architectures and statistical machine learning tasks is an exciting new area beginning to emerge, which will require ideas such as those addressed in this manuscript to facilitate data sourcing.

As noticed in Murakami et al. (2016), such sensing platforms can also be practically useful to exploit the growth of mobile smartphone ownership and tap into smartphone users to contribute sensing data that are easily obtained via the available sensing capabilities of their smartphones. This allows the sensing platform to then utilise this spatially distributed data to perform important practical tasks such as to estimate some statistics of a spatial event or to conduct spatial regression. Other examples of mobile crowd sensing applications for spatial monitoring include environmental temperature monitoring Overeemn et al. (2013), traffic monitoring Thiagarajan et al. (2009) or earthquake detection in early warn-
Privacy Aware Participatory Sensing Mechanisms

Figure 1: Privacy model of users: privacy-concerned users can declare a cloaking region that contains their true location instead of providing fine-grained location information to the crowdsourcer.

ing systems Minson et al. (2015). All of these aforementioned application domains have benefited from the fact that mobile crowdsourced data has improved the spatial coverage of the collected dataset and therefore improved the resolution of spatial regression model estimations. In our framework, we consider the aspect of spatial coverage as one of the main objectives of an incentive mechanism used by spatial monitoring applications.

Additionally, current privacy-preserving works such as Nissim et al. (2012); Yang et al. (2013); Singla and Krause (2013) have attempted to address the user location privacy problem in the crowd sensing domain. This is because the privacy issues can easily deter potential users from participating, which in turn reduces both the amount of available data as well as spatial coverage of the data used by the crowdsourcer.

1.1 Location-Privacy Preservation and Incentivised Crowdsourced Data

Existing privacy-preserving works that offer location privacy via location or data perturbation are not directly applicable to crowd sensing applications that require specific and true locations. For example, it would be unacceptable for a traffic monitoring application if there was a traffic congestion in road X, but due to location or data perturbation, another road Y or a non-congested status was reported respectively. Thus, it is vital for incentive models to address the spatial coverage and location privacy issues concurrently.

The most common incentive mechanism is the auction game, see examples in Yang et al. (2012); Nissim et al. (2012); Singla and Krause (2013) and Restuccia et al. (2016). In such settings the smartphone workers submit bids (e.g., prices or efforts in Luo et al. (2016b)) for their sensing data while the crowdsourcer selects the set of workers with the lowest bids and pays them accordingly.

An alternative and fundamentally different game-theoretic approach to model the incentive problem is the Stackelberg leader-follower game where the crowdsourcer (leader) first decides on the total reward to pay workers (i.e., the leader’s budget constraint) while the workers (followers) then individually decide on the amount of data to contribute, see Yang et al. (2012); Luo et al. (2016a).

However, different from the system models used in Yang et al. (2012); Luo et al. (2016a) where the crowdsourcer first informs the workers of the total reward offered. Then subsequently, the workers compute the optimal data to contribute (which may not be intuitive or practically feasible for the smartphone users in practice), we allow the crowdsourcer to specify the exact reward offered to each individual worker (see stage 2 in Fig. 2). In this way, the workers can be assured that the offered rewards are always their optimal (Nash equilibrium) solution and there is no need to share all the information on the workers’ costs and privacy preferences between other competing workers.
Privacy-aware incentive mechanisms were considered in Nissim et al. (2012); Yang et al. (2013); Singla and Krause (2013) which used data perturbation or dummy locations Kido et al. (2005) to protect location privacy. However, the methods may not be applicable to many spatial monitoring applications where a dataset with perturbed data or dummy location may trigger a false alarm and make the application unreliable. In this paper, we allow workers to obfuscate their precise location information by declaring a coarse-grained region (see Fig. 1) that encompasses their true location (similar to the location cloaking principle in ??). This step is important as privacy concerns can deter potential workers from participating in the crowd sensing activity. In addition, the work in ? found that smartphone users were more willing to provide coarse-grained location information than fine-grained information. Hence, our Stackelberg model may still buy data (albeit with lower quality location information) from privacy-sensitive workers to improve the spatial coverage of the collected dataset. Furthermore, the cloaking region technique is more practical for real-world applications that require reliable information. Although the location information of the workers may be imprecise due to the location cloaking, but it is still accurate as there is no data perturbation or dummy locations involved.

Additionally, existing Stackelberg models assume that workers sell undifferentiated goods (data) and do not consider the quality (e.g., granularity) of each worker’s data and the workers’ contribution to the spatial coverage of the collected dataset. However, adding location information increases the computational complexity of the problem and thus, Feng et al. (2014) proposed an auction-based approximation algorithm to assign the workers’ sensing tasks. The authors assumed that the sensing platform periodically publishes sensing tasks for specific locations of interest and did not explicitly address the issue of improving the spatial coverage of the collected dataset in general.

1.2 Our Contributions

We demonstrate that an appropriate incentive mechanism to model the hierarchical relationship between the crowdsourcer and the smartphone users is the Stackelberg (leader-follower) game model used in Yang et al. (2012); Luo et al. (2016a). In the Stackelberg model, the crowdsourcer (leader) commits a reward strategy that is observed by the smartphone users (followers) who then strategize the amount of data to sell. However, existing Stackelberg incentive models simply select user data independently of their physical location Feng et al. (2014) and do not attempt to improve the spatial coverage of the dataset.

Therefore, we propose extending the existing Stackelberg incentive models to include the privacy-awareness property as well as to improve the spatial coverage of the collected dataset. Our model allows privacy-sensitive users to submit coarse-grained (or quantized) location information which could still be useful to the crowdsourcer. We then study the properties of the proposed Stackelberg incentive model analytically and present efficient algorithmic solutions. Our proposed Stackelberg incentive model does not require a trusted third party for privacy and can protect users against a crowdsourcer who cannot be trusted to anonymize the smartphone users’ location information. The current work extends our preliminary work in Koh et al. (2017), by accommodating bounds in the workers’ sensing data, analyzing the existence, uniqueness and Pareto efficiency of the Stackelberg equilibrium, and conducting a real-world sensing case study to demonstrate the practicality of the proposed solution.

The main contributions of the paper are as follows:

- We propose a novel spatial privacy-aware Stackelberg incentive scheme that allows privacy-sensitive mobile smartphone users to quantize their location information using
cloaking regions. Our proposed solution also seeks to improve the spatial coverage of the collected dataset.

- We prove that under our Stackelberg incentive mechanism, we are able to develop a unique Nash equilibrium for the Followers game, and obtain a closed-form expression for each mobile smartphone user’s optimal data contribution.

- We also prove the existence of a Stackelberg equilibrium when the crowdsourcer (leader) imposes constraints on the minimum and maximum amount of data contribution from each user and we further derive sufficient conditions for the Stackelberg equilibrium to be Pareto efficient.

- We demonstrate via simulations using a real-world sensing dataset that the proposed spatial privacy-preserving Stackelberg game produces a mechanism that for a common cost, will produce greater spatial diversity in crowdsourced data, leading to better model predictive performance in applications such as spatial field reconstruction compared to two other coverage-maximizing schemes.

2. Problem Formulation and Analysis

In this paper, we address the following problem statement: suppose there is a crowdsourcer (who is the data buyer and leader on the demand side) who aims to buy sensing data from mobile smartphone users (who are the workers and followers on the supply side) for applications such as spatial monitoring; we aim to design an incentive mechanism such that the collected dataset: (i) has good spatial coverage, and (ii) is location privacy-preserving for the workers. We first present our system model and the proposed Stackelberg incentive model before studying the proposed game analytically. We extend the results in (Yang et al., 2012. Section 3) to take into account the workers’ location granularities and regions. The table of notation is given in Table 1.

| Table 1: Notation |
|-------------------|
| $c_i$ | Sensing cost incurred per unit of data by worker $i$ where $c_i \in (0, \bar{c}]$. |
| $c_{-i}$ | Sensing cost incurred per unit of data by all workers other than worker $i$. |
| $\rho_i$ | Location granularity of worker $i$ where $\rho_i \in [\rho, \bar{\rho}]$. |
| $\rho_{-i}$ | Location granularities of all workers other than worker $i$. |
| $l_i, l'_i$ | Partitioned region and cloaking region of worker $i$ respectively where $l'_i \in l_i$. |
| $t_i$ | Amount of data contributed by worker $i$. |
| $t_{-i}$ | Amount of data contributed by all workers other than worker $i$. |
| $R_l$ | Sum of rewards allocated to workers in region $l$ (similarly, $R_{l_i}$ refers to the sum of rewards allocated to worker $i$’s region $l_i$). |
| $\alpha_i$ | System parameter for crowdsourcer. |
| $\beta$ | Diminishing returns parameter for crowdsourcer. |
| $U_{CS}, u_i$ | Utility function of crowdsourcer and worker $i$ respectively. |
| $\mathcal{I}$ | Set of all workers. |
| $N_l$ | Number of workers in region $l$. |
| $Q_l$ | Set of participating workers in region $l$ (similarly, $Q_{l_i}$ refers to the set of participating workers in the region $l_i$ where worker $i$ is located). |
| $\mathcal{J}_l$ | Set of participating workers $i$ in region $l$ with $t_i = \bar{t}$. |
| $\mathcal{K}_l$ | Set of participating workers $i$ in region $l$ with $t_i < \bar{t}$. |
2.1 Model and Assumptions

We consider a crowd sensing system that consists of a set of $\mathcal{I} = \{1, \ldots, N\}$ workers and a single crowdsourcer who partitions the entire spatial area of interest into a set of $L$ regions denoted by $\mathcal{L}$. We assume that the workers are rational and non-cooperative, i.e., each worker maximizes its own utility. Each worker $i \in \mathcal{I}$ has its own sensing cost per unit of data $c_i \in (0, \bar{c}]$, location granularity $\rho_i \in [\underline{\rho}, \bar{\rho}]$, cloaking region $l'_i$ (which may be of different granularity for each worker) and its corresponding region $l_i$ (defined by the crowdsourcer) where $l'_i \subseteq l_i \subseteq \mathcal{L}$. The interaction model between the crowdsourcer and the mobile smartphone users is illustrated in Fig. 2. Essentially, the crowdsourcer collects the worker profiles ($c_i, \rho_i, l_i$) and selects the optimal set of workers that maximizes its utility. The crowdsourcer then offers the selected workers a reward in exchange for some specified amount of sensing data. Next, the selected workers (which we refer to as participating workers) will proceed to collect their sensing data and transmit it along with their cloaking region $l'_i$ to the crowdsourcer and receive their rewards in return. Note that there are two types of regions $l_i$ and $l'_i$ in our model: $l_i$ is the initial coarse-grained partitioned region defined by the crowdsourcer and $l'_i$ is worker $i$'s cloaking region, which is submitted only when the worker is selected.

We assume that the granularity of the worker $i$'s submitted cloaking region $l'_i$ is proportional to its location granularity parameter $\rho_i$, since a privacy-sensitive worker (with low $\rho_i$) is likely to provide only coarse-grained location information, while a privacy-insensitive worker (with high $\rho_i$) is more likely to provide finer-grained location information. Hence, the parameter $\rho_i$ allows the crowdsourcer to differentiate between workers with different levels of sensitivity to privacy and preserve the privacy of unselected workers since they only reveal more information on their locations when they are selected. Note that the workers can anonymize their precise location information via cloaking regions with area inversely proportional to $\rho_i$ and do not rely on the crowdsourcer for anonymization. In practice, this can be implemented in the crowdsourcing software on the workers’ smartphone devices to allow each worker to select cloaking regions with different (possibly discrete) location granularities.

We model the incentive mechanism as a Stackelberg game, which consists of the crowdsourcer (data buyer) as the leader and the $N$ smartphone users (workers and data contributors) as the followers. The crowdsourcer acts first and commits a reward strategy.
while the workers subsequently choose their best responses after observing the crowdsourcer’s strategy. The strategy of the crowdsourcer is the reward for each partitioned region $R = (R_1, \ldots, R_L)$ and the strategy of worker $i$ is the amount of data contributed $t_i \geq 0$. The crowdsourcer only optimizes the reward $R_l$ allocated to each partitioned region $l \in L$ and subsequently offers each participating worker $i$ in the partitioned region $l$ a fraction of $R_l$ depending on the proportion of data contribution, weighted by the location granularity:

$$\text{Worker } i\text{'s offered reward} = \frac{t_i \rho_i}{\sum_{j \in Q_{li}} t_j \rho_j} R_{li},$$

(1)

where $Q_{li}$ is the set of participating workers $i$ in partitioned region $l_i \in L$ with $t_i > 0$ and we assume $|Q_{li}| > 1$. A similar reward function has also been used in Yang et al. (2012).

2.2 Utility Functions for Crowdsourcer and Workers

In this section we introduce the formulations for the utility functions for each party in the crowdsourcing data mechanism.

Crowdsourcer: We define the utility function of the crowdsourcer to be $U_{CS}(R; t) = \sum_{i \in I} f_d(t_i, l_i, \rho_i)$, where $f_d(t_i, l_i, \rho_i)$ is a function of worker $i$’s amount of data contributed $t_i$, partitioned region $l_i$, and its location granularity $\rho_i$. For simplicity, we let $f_d(t_i, l_i, \rho_i) = \alpha_i t_i \beta$ where $\alpha_i > 0$ is a function of $(l_i, \rho_i)$, and the parameter $0 < \beta < 1$ is used to control the rate of diminishing returns on each worker’s data. The power function was also used in Powell and Batt (2008) to model diminishing returns. Thus, the crowdsourcer’s utility function is given by:

$$U_{CS}(R; t) = \sum_{i \in I} \alpha_i t_i \beta.$$  

(2)

To increase the coverage area of the collected dataset, the crowdsourcer can assign a higher $\alpha_i$ value to workers located at less populated regions $l_i$. In addition, a higher $\alpha_i$ value can be assigned to workers that provide finer location granularity $\rho_i$. The $\alpha_i$ parameter allows the crowdsourcer to differentiate between the quality (e.g., spatial coverage area and the granularity of the location information) of each worker’s data.

Workers: We define the utility function of worker $i$ to be the amount of reward received from the crowdsourcer as defined in (1) minus the cost incurred for obtaining the data:

$$u_i(t_i; t_{-i}, R_{li}) = \frac{t_i \rho_i}{\sum_{j \in Q_{li}} t_j \rho_j} R_{li} - c_i t_i,$$

(3)

where $t_{-i}$ is a vector of the amount of data contributed by all workers except worker $i$.

2.3 Stackelberg Game Formulation

Given that the crowdsourcer wants to achieve a large coverage area for its dataset while satisfying a budget constraint $R^{\text{budget}}$, a minimum amount of reward $R^{\text{min}}_l > 0$ for each region $l$ (this allows the crowdsourcer to specify more important regions), and a maximum amount of reward allocation $R^{\text{max}}_l > 0$ for each region $l$ (which may be set to $\infty$), it solves the following optimization problem:
Problem 1 ($\alpha, \beta, R_{\text{min}}, R_{\text{max}}, R_{\text{budget}}$):

\[
\begin{align*}
\text{maximize} & \quad \sum_{i \in I} \alpha_i t_i \beta \\
\text{subject to} & \quad R_{\text{min}} l \leq R_l \leq R_{\text{max}} l, \quad \forall l \in L, \\
& \quad \sum_{l \in L} R_l \leq R_{\text{budget}}, \quad (4)
\end{align*}
\]

where $t_i$ is the optimal solution to Problem 2.

Each worker $i$ solves the following optimization problem:

Problem 2 ($i, t_{-i}, \rho_i, \rho_{-i}, R_l, c_i$):

\[
\begin{align*}
\text{maximize} & \quad \frac{t_i \rho_i}{\sum_{j \in Q_l} t_j \rho_j} R_l - c_i t_i, \\
\text{subject to} & \quad t_i \geq 0. \quad (5)
\end{align*}
\]

Problems 1 and 2 form the Stackelberg game and our goal is to find the Stackelberg equilibrium point(s) where neither the crowdsourcer nor the workers have incentive to deviate. A Stackelberg equilibrium (see Definition 1) is a subgame-perfect Nash equilibrium such that no player can improve its utility by unilaterally deviating from its strategy.

**Definition 1** (Stackelberg Equilibrium). Let $R^*$ be the optimal solution for the crowdsourcer, obtained by solving Problem 1, and $t^*$ be the optimal solution for the workers, obtained by solving Problem 2. The strategy profile $(R^*, t^*)$ is a Stackelberg equilibrium for the proposed Stackelberg incentive model if the following conditions are satisfied for any $(R, t)$ where:

\[
\begin{align*}
U_{CS}(R^*; t^*) & \geq U_{CS}(R; t^*), \\
u_i(t_i^*; t_{-i}^*, R^*) & \geq u_i(t_i; t_{-i}^*, R^*), \quad \forall i \in I.
\end{align*}
\]

We apply the backward induction method to analyze the proposed Stackelberg incentive model. First, we start with the Followers game (a non-cooperative game played by all the workers) and study the predicted best response $t_i^*$ (solution of Problem 2) for each worker $i$ as a function of the reward $R_l$ offered by the crowdsourcer and the strategies of the other workers $t_{-i}$. Subsequently, we analyze the best response of the crowdsourcer in Problem 1.

**2.4 Nash Equilibrium of Followers game**

We first consider the Followers game given by the triplet $(I, \{t_i\}_{i \in I}, \{u_i\}_{i \in I})$ where $I$ is the player set of $N$ workers and $u_i$ is the utility function of worker $i$. By Lemma 2, there exists a unique Nash equilibrium point in the Followers game. Next, we derive the unique Nash equilibrium of the Followers game in Theorem 3.

**Lemma 2.** A unique Nash equilibrium exists in the Followers game $(I, \{t_i\}_{i \in I}, \{u_i\}_{i \in I})$.

**Proof:** See Appendix .1.
Theorem 3. The Followers game \((\mathcal{I}, \{t_i\}_{i\in\mathcal{I}}, \{u_i\}_{i\in\mathcal{I}})\) has a unique Nash equilibrium given by the following closed-form expression:

\[
t^*_i = \begin{cases} 
\frac{1}{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}} \left(1 - \frac{(|\mathcal{Q}_{l_i}| - 1) \frac{c_i}{\rho_i}}{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}}\right) R_{l_i} / \rho_i & \text{if } i \in \mathcal{Q}_{l_i}, \\
0 & \text{otherwise,}
\end{cases}
\]

(6)

where \(\mathcal{Q}_{l_i}\) is the set of participating workers in region \(l_i\).

Proof: See Appendix .2.

We say that worker \(i\) is a participating worker if \(t^*_i > 0\). Hence, from (6), all participating workers \(i \in \mathcal{Q}_{l_i}\) should satisfy the constraint:

\[
\frac{|\mathcal{Q}_{l_i}| - 1}{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}} \left(1 - \frac{(|\mathcal{Q}_{l_i}| - 1) \frac{c_i}{\rho_i}}{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}}\right) R_{l_i} / \rho_i > 0
\]

\[
\iff \left(1 - \frac{(|\mathcal{Q}_{l_i}| - 1) \frac{c_i}{\rho_i}}{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}}\right) > 0
\]

\[
\iff \frac{c_i}{\rho_i} < \frac{\sum_{j\in\mathcal{Q}_{l_i}} \frac{c_j}{\rho_j}}{|\mathcal{Q}_{l_i}| - 1}.
\]

(7)

Subsequently, we will make use of this constraint in Algorithm 1 (modified from (Yang et al., 2012, Algorithm 1)), which computes the Nash equilibrium solution for the Followers game.

The optimal \(t^*_i\) for the workers is given by the Nash equilibrium solution (6) of the Followers game. However, the expression in (6) requires knowledge of the set of participating workers \(\mathcal{Q}_{l_i}\), the sensing costs \(c_i\), and the location granularity \(\rho_i\) of all workers \(i \in \mathcal{I}\). Hence, we propose Algorithm 1 which makes use of (7) to greedily compute \(\mathcal{Q}_{l_i}\) and solve for the optimal \(t^*_i\) of the workers (in terms of \(R_{l_i}\)). The values of \(t^*_i\) can then be substituted into Problem 1 to solve for \(R^*\).

Lemma 4. Assume that there are at least two workers in each region \(l\). Algorithm 1 selects the set of participating workers \(\mathcal{Q}_{l}\) that achieves the unique Nash equilibrium solution of the Followers game.

Proof: See Appendix .3.

Theorem 5 proves the correctness of Algorithm 1 while Lemma 4 simply states that the \(\mathcal{Q}_{l}\) set computed in Algorithm 1 is correct.

Theorem 5. Assuming that there are at least two workers in each region \(l\), and the set of \(\mathcal{Q}_{l}\) from Lemma 4, then Algorithm 1 outputs the unique Nash equilibrium solution of the Followers game.

Proof: See Appendix .4.

2.5 Stackelberg Equilibrium

Using the analytical result (6) for the Followers game, the crowdsourcer can optimize its reward strategy \(R\) efficiently by substituting the analytical result into its utility function in (2) to obtain:

\[
U_{CS}(R; t) = \sum_{i\in\mathcal{I}} \alpha_i (\tau_i R_{l_i})^\beta,
\]

(8)
where \( \tau_i = \frac{(|Q_i| - 1)}{\rho_i \sum_{j \in Q_i} \frac{r_j}{|Q_i|}} \left( 1 - \frac{(|Q_i| - 1)}{\rho_i \sum_{j \in Q_i} \frac{r_j}{|Q_i|}} \right) \).

Although it is not trivial to obtain a closed-form expression for \( \mathbf{R} \) that maximizes (8) while satisfying the constraints in (4), we show in Theorem 6 that there exists a unique Stackelberg equilibrium which results in a stable equilibrium strategy profile. This allows the crowdsourcer to uniquely predict the behaviors of the workers and efficiently compute its optimal \( \mathbf{R}^* \). Both Theorems 3 and 6 extend (Yang et al., 2012, Theorem 2) to the case where the workers’ location granularities \( \rho \) and regions \( l \) are also considered.

**Theorem 6.** The proposed Stackelberg incentive model has a unique Stackelberg equilibrium.

**Proof.** Recall from Theorem 3 that the Followers game has a unique Nash equilibrium point. It can be easily shown that the best strategy set of the crowdsourcer is convex and compact since the domain of \( \mathbf{R} \) is a Cartesian product of closed intervals, and \( U_{CS} \) is continuous in \( \mathbf{R} \). Hence, we need to show the strict concavity of \( U_{CS} \) to conclude that there exists a unique maximum point. The partial derivatives of \( U_{CS} \) with respect to \( R_l \) are as follows:

\[
\begin{align*}
\frac{\partial U_{CS}}{\partial R_l} &= \sum_{i,l_i=l} \alpha_i \beta (\tau_i R_l)^{\beta - 1}, \\
\frac{\partial^2 U_{CS}}{\partial R_l^2} &= \sum_{i,l_i=l} \alpha_i (\beta - 1) \beta (\tau_i R_l)^{\beta - 2}, \quad \frac{\partial^2 U_{CS}}{\partial R_l \partial R_k} = 0,
\end{align*}
\]

where \( \tau_i \) is given in (8). Note that \( \frac{\partial^2 U_{CS}}{\partial R_l^2} < 0 \) as we assume \( \alpha_i > 0 \) for all \( i \in \mathcal{I} \) and \( 0 < \beta < 1 \). Since the Hessian matrix of \( U_{CS} \) is a diagonal matrix, its eigenvalues are given by \( \frac{\partial^2 U_{CS}}{\partial R_l^2} \), which are all strictly negative. This implies that the Hessian matrix is negative definite and thus, \( U_{CS} \) is strictly concave in \( \mathbf{R} \).

Using the results from Theorem 5 (which gives the closed-form solution to Problem 2), the Stackelberg equilibrium solution can be computed by solving Problem 1, which is a convex optimization problem. By Theorem 6, there exists an unique reward strategy \( \mathbf{R}^* \) that maximizes the crowdsourcer’s utility in Problem 1. This means that \( \mathbf{R}^* \) can be efficiently computed using well-known interior point methods.

3. Analysis of Optimal Declared Sensing Costs

In this section, we study the strategies of the workers in selecting their optimal sensing costs to declare to the crowdsourcer. This is essential as each worker may declare a false sensing cost to maximize its utility. For this section, we let the declared sensing cost for each worker \( i \) be denoted by \( c_i' \) and its true sensing cost be denoted by \( c_i \). Note that the true sensing cost \( c_i \) is unknown to the crowdsourcer and other workers \( j \in Q_i \) where \( j \neq i \).

Suppose that each worker \( i \) may lie about its sensing cost to maximize its utility \( u_i \), in such a dishonest reporting scenario, the optimization problem previously specified in Problem 2 is replaced with that considered in Problem 3 below. The interaction model previously shown in Fig. 2 needs to be updated as each worker \( i \) requires additional information on \( Q_i \) and \( \sum_{j \in Q_i} \frac{r_j}{\rho_j} \) to solve Problem 3. One way for the workers to obtain the required information would be for the crowdsourcer to provide a platform for the workers to view.
the required information for a fixed time period. If the workers can declare or modify their $c'_i$ values during the fixed time period, then they can achieve the Nash equilibrium solution given in Theorem 9.

Problem 3 ($i, c_i, c_{-i}, R_i, R_i'$):

\[
\begin{align*}
\text{maximize} & \quad \sum_{j \in Q_i} t_j \rho_j - c_i t_i, \\
\text{subject to} & \quad 0 \leq c'_i \leq \bar{c},
\end{align*}
\]

where $t_i = \frac{|Q_i|-1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}} \left(1 - \frac{|Q_i| - 1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}\right) R_i$ if $i \in Q_i$, and 0 otherwise, according to (6).

We first derive the expression for the worker $i$’s utility in terms of its declared sensing cost and the declared sensing costs of the other participating workers $j$ in its region $l_i$. Using Algorithm 1, the crowdsourcer offers to buy $t_i = \frac{|Q_i|-1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}} \left(1 - \frac{|Q_i| - 1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}\right) R_i$ amount of data from each worker $i$ according to (6). From (26), we have the expression:

\[
\sum_{j \in Q_i} t_j \rho_j = \frac{|Q_i|-1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}} R_i.
\]

We substitute the expressions for $t_i$ and $\sum_{j \in Q_i} t_j \rho_j$ into the utility of worker $i \in Q_i$ in (3) to obtain:

\[
u_i(t_i; t_{-i}, R_i) = \frac{|Q_i|-1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}} R_i \rho_i - c_i \left(1 - \frac{|Q_i| - 1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}\right) R_i = \left(1 - \frac{|Q_i| - 1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}\right) \left(1 - \frac{|Q_i| - 1}{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}\right) R_i.
\]

Next, we have the following Theorem 9 on the Nash equilibrium solution of the Followers game when the workers can optimize their declared sensing costs.

Lemma 7. Suppose that worker $i \in Q_i$ declares a sensing cost $c'_i < c_i$, then it always incurs a negative utility $u_i < 0$ if $\frac{c_i}{\rho_i} > \frac{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}{|Q_i|-1}$ and gets a positive utility $u_i > 0$ if $\frac{c_i}{\rho_i} < \frac{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}{|Q_i|-1}$.

Proof. Suppose worker $i$ is not in the set of participating workers $Q_i$ if it does not lie about its sensing cost, i.e., $\frac{c_i}{\rho_i} \geq \frac{\sum_{j \in Q_i} \frac{c_j}{\rho_j}}{|Q_i|-1}$. To be in the set of participating workers $Q_l$, the worker
i can declare a sensing cost $c'_i < c_i$ such that $\frac{c'_i}{\rho_i} < \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$. In order for its utility to be negative, i.e., $u_i < 0$, we have the following necessary condition from (11): $\frac{c_i}{\rho_i} > \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$.

Similarly, for $u_i > 0$, we have the following necessary condition from (11): $\frac{c_i}{\rho_i} < \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$. In order for its utility to be negative, i.e., $u_i < 0$, we have the following necessary condition from (11): $\frac{c_i}{\rho_i} > \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$. Similarly, for $u_i > 0$, we have the following necessary condition from (11): $\frac{c_i}{\rho_i} < \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$. Hence, the proof is complete.

Lemma 8. Given the set of participating workers $Q_l$, the worker $i \in Q_l$ can maximize its utility $u_i$ if it declares a sensing cost of $c'_i = \frac{\left|Q_l\right|}{\left|Q_l|-2}\left(\frac{\rho_i}{\rho_i} - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}\right)^2 - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}$, $\forall i \in Q_l$.

Proof: See Appendix .5.

Theorem 9. Assuming that each worker $i \in Q_l$ optimizes its sensing cost $c'_i$ using Lemma 8, then the Nash equilibrium solution of the Followers game occurs when

$$c'_i = \frac{\left|Q_l\right|}{\left|Q_l|-2}\left(\frac{\rho_i}{\rho_i} - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}\right)^2 - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}, \forall i \in Q_l. \tag{12}$$

Proof. By Lemma 7, we have the constraint $\frac{c_i}{\rho_i} < \frac{\sum_{j \in Q_l, j \neq i} c'_j}{|Q_l|-1}$ for all workers $i \in Q_l$ to ensure that all participating workers $i \in Q_l$ have a positive utility, i.e., $u_i > 0$. By Lemma 8, the worker $i \in Q_l$ can maximize its utility $u_i$ if it declares a sensing cost of

$$c'_i = \frac{\left|Q_l\right|}{\left|Q_l|-2}\left(\frac{\rho_i}{\rho_i} - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}\right)^2 - \frac{\sum_{j \in Q_l, j \neq i} c'_j}{\rho_j}, \forall i \in Q_l.$$

Remarks: If we assume that the workers $i \in Q_l$ have a fixed time period to declare or modify their $c'_i$ values and that the crowdsourcer provides a platform for each individual worker to observe the total declared sensing cost $c'_j$ and location granularities $\rho_j$ of all other workers $j \in Q_l$ in the same region, then the equilibrium point in (12) can be achieved.
4. Adding Bounds on the Amount of Workers’ Data and Achieving Pareto Efficiency

In this section, we study how the crowdsourcer is able to introduce bounds on the workers’ data contributions and how it can achieve Pareto efficiency for the Stackelberg equilibrium.

4.1 Bounds on the Amount of Contributed Data $t_i$

In a practical real-world crowdsourcing application, the crowdsourcer may impose (lower and upper) bounds on the amount of data contribution from each worker $i$ due to various reasons. For example, it may not be useful to the crowdsourcer if each worker only contributes a small amount of data, hence we consider the imposition of a lower bound $t_\rho$ on the workers’ data contribution. In the other extreme, to maximize the spatial coverage of the obtained data, it may not be useful to the crowdsourcer if there are a small number of workers who are monopolies. Hence, we also consider the imposition of an upper bound $\overline{t_\rho}$ on the workers’ data contribution. For simplicity, we assume that the same bounds $t_\rho$ and $\overline{t_\rho}$ apply to all participating workers.

Next, we have the following Theorem 10 to introduce the lower and upper bounds on the worker $i$’s $t_\rho$ values and Theorem 14, which states the sufficient condition for the bounds. Finally, we present Algorithm 2, which computes the bounded Nash equilibrium solution for the Followers game.

Theorem 10. Assume that $\frac{c_i}{\rho_i} \leq 1$ for all $i \in \mathcal{I}$. The crowdsourcer is able to introduce both the lower bound $t_\rho$ and upper bound $\overline{t_\rho}$ in the Stackelberg equilibrium of all participating workers $i$ in region $l$ if the following constraints are satisfied:

$$
\frac{c_i}{\rho_i} \leq \frac{\sum_{j \in \mathcal{Q}_l} \frac{c_j}{\rho_j}}{|\mathcal{Q}_l| - 1} \left(1 - \frac{t_\rho \sum_{j \in \mathcal{Q}_l} \frac{c_j}{\rho_j}}{(|\mathcal{Q}_l| - 1)R_l}\right), \quad \forall i \in \mathcal{Q}_l,
$$

$$
R_l^{\max} = \overline{t_\rho} \left[\left(1 - \frac{(|\mathcal{Q}_l| - 1)\frac{c_M}{\rho_M}}{\sum_{j \in \mathcal{Q}_l} \frac{c_j}{\rho_j}}\right)^{-1}\right],
$$

where we let the sensing cost and location granularity of the worker $i$ with the least $\frac{c_i}{\rho_i}$ value in region $l$ be denoted by $c_M$ and $\rho_M$ respectively.

Proof. We first show the following two Lemmas, which are used to prove the theorem:

Lemma 11. Suppose that $\frac{c_i}{\rho_i} \leq 1$ for all $i \in \mathcal{I}$ and (14) is true. Then the crowdsourcer is able to introduce a lower bound $t_\rho$ in the Stackelberg equilibrium of all participating workers $i$ in region $l$.

$$
\frac{c_i}{\rho_i} \leq \frac{\sum_{j \in \mathcal{Q}_l} \frac{c_j}{\rho_j}}{|\mathcal{Q}_l| - 1} \left(1 - \frac{t_\rho \sum_{j \in \mathcal{Q}_l} \frac{c_j}{\rho_j}}{(|\mathcal{Q}_l| - 1)R_l}\right), \quad \forall i \in \mathcal{Q}_l,
$$

We first prove Lemma 11. To show that the Stackelberg solution exists in the constrained region, one must show that $t^*_i \rho_i \geq t_\rho$ for all $i \in \mathcal{Q}_l$. To prove that $t^*_i \rho_i \geq t_\rho$ for all $i \in \mathcal{Q}_l$, 

we first let \( \frac{c_i}{\rho_i} = \sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j} - 1 \) where \( \delta_i \geq 0 \). We take the solution in (6) and substitute the \( \frac{c_i}{\rho_i} \) expression from above and multiply both sides by \( \rho_i \) to obtain:

\[
t^*_i \rho_i = \frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}} \left( \left(\frac{|Q_i|}{|Q_i| - 1}\right) \left(1 - \frac{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}}{|Q_i| - 1} \right) - \delta_i \right) R_i
\]

\[
= \frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}} \left(1 - \frac{t \rho \sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}}{|Q_i| - 1} \right) R_i
\]

\[
= t \rho + \left(\frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}}\right)^2 R_i \delta_i. \tag{15}
\]

Since \( \delta_i \geq 0 \), we conclude from (15) that \( t^*_i \rho_i \geq t \rho \) for all \( i \in \mathcal{Q}_i \).

To verify that the introduction of the lower bound preserves the concavity of the crowd-sourcer’s Problem 1, we analyze the Hessian matrix of \( U_{CS} \) with respect to \( R_i \).

We first apply the chain rule: \( \frac{\partial^2 U_{CS}}{\partial R_i^2} = \frac{\partial^2 U_{CS}}{\partial t^2} \times \frac{\partial t}{\partial R_i} \). From (2), we obtain \( \frac{\partial^2 U_{CS}}{\partial t^2} = \beta(\beta - 1) \sum_{i \in \mathcal{I}} \alpha_i t_i \beta - 2 \), where \( \beta < 1 \). Hence, we have \( \frac{\partial^2 U_{CS}}{\partial t^2} < 0 \). We now attempt to derive \( \frac{\partial t_i}{\partial R_i} \). From (15), we have \( t_i \rho_i = t \rho + \left(\frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}}\right)^2 R_i \). If we assume that \( \frac{c_i}{\rho_i} \leq 1 \) for all \( i \in \mathcal{I} \), then the term \( \frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}} \) is inversely proportional to \( R_i \). This is because \( |Q_i| \propto R_i \) according to (14) and the term \( \sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j} - \delta_i \) is proportional to \( |Q_i| \) but increases at a rate slower than \( |Q_i| \) when \( \frac{c_i}{\rho_i} \leq 1 \). Therefore, we conclude from (15), that \( t_i \) is a monotonically non-decreasing function of \( R_i \). This implies convexity, i.e., \( \frac{\partial^2 t_i}{\partial R_i^2} \geq 0 \) given that \( t_i \) is continuous. Since \( \frac{\partial^2 U_{CS}}{\partial t^2} < 0 \) and \( \frac{\partial t_i}{\partial R_i} \geq 0 \), we have \( \frac{\partial^2 U_{CS}}{\partial R_i^2} \leq 0 \). In (9), it was shown that \( \frac{\partial^2 U_{CS}}{\partial R_i \partial R_k} = 0 \). Since \( \frac{\partial^2 U_{CS}}{\partial R_i^2} \leq 0 \), the Hessian matrix of \( U_{CS} \) is negative semidefinite for all \( R_i \in \mathbb{R} \). Thus, we conclude that \( U_{CS} \) is concave in \( R \).

To introduce the lower bound \( t \rho \) constraint in the proposed Stackelberg incentive model, the constraint (14) should be used instead of constraint (7) in line 6 of Algorithm 1.

Next, we derive the sufficient conditions for the upper bound \( \overline{t} \rho \) on the workers’ data contributions. For each region \( l \), we let the sensing cost and location granularity of the “cheapest” worker \( i \) with the least \( \frac{c_i}{\rho_i} \) value be denoted by \( c_i^{M} \) and \( \rho_i^{M} \) respectively. Using the \( t^*_i \) expression in (6), we proceed to derive the maximum \( t_i \rho_i \) contributed by the “cheapest” worker in each region \( l \):

\[
t^{M}_i \rho^{M}_i = \frac{|Q_i| - 1}{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}} \left(1 - \frac{\sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j}}{|Q_i| - 1} \right) R_i. \tag{16}
\]

This leads to the following Lemma 12.
Lemma 12. The crowdsourcer is able to introduce an upper bound $\overline{t\rho}$ in the Stackelberg equilibrium of all participating workers $i$ in region $l$ if:

$$R_{l}^{\text{max}} \leq \overline{t\rho} \left[ \left| Q_{l} \right| - 1 \left( 1 - \frac{\left( \left| Q_{l} \right| - 1 \right) c_{i}^{M} \rho_{i}^{M}}{\sum_{j \in Q_{l}} c_{j} \rho_{j}} \right) \right]^{-1}. \tag{17}$$

We now prove Lemma 12. According to (6), the worker $i \in Q_{l}$ with the least $\frac{c_{i}}{\rho_{i}}$ value will contribute the most amount of $t_{i} \rho_{i}$ in its region $l$. Thus, to prove that the constraint in (17) leads to $t_{i} \rho_{i} \leq \overline{t\rho}$ for all $i \in Q_{l}$, it is sufficient to show that the data contribution $t_{i}^{M} \rho_{i}^{M}$ from the participating worker $i$ with the least $\frac{c_{i}}{\rho_{i}}$ value is less than or equal to the upper bound $\overline{t\rho}$, i.e., $t_{i}^{M} \rho_{i}^{M} \leq \overline{t\rho}$. Indeed, we substitute the $R_{l}^{\text{max}}$ term from (17) into (16) to obtain $t_{i}^{M} \rho_{i}^{M} = \overline{t\rho}$. This implies that $t_{i} \rho_{i} \leq \overline{t\rho}$ for all $i \in Q_{l}$. Hence, the proof is complete. Note that the constraint on $R_{l}$ simply constraints the feasible region and does not affect the concavity of the crowdsourcer’s Problem 1.

Corollary 13. The Stackelberg equilibrium can be shown to exist under the constraints in Theorem 10 but the uniqueness property of the unconstrained solution is lost. Furthermore, the solution can be obtained via the same algorithmic solution as utilized in the original unconstrained case, with a minor modification to Algorithm 1 as shown in Algorithm 2.

Finally, we have the following Theorem 14, which states the sufficient conditions where each worker’s optimal data contribution satisfies the crowdsourcer’s $t\rho$ and $\overline{t\rho}$ bounds. This allows the crowdsourcer to estimate the number of participating workers in a region $l$ given its chosen bounds $t\rho$ and $\overline{t\rho}$, and reward allocation $R_{l}$.

Theorem 14. Suppose (18) is true for all $i \in Q_{l}$. Then the worker $i$’s optimal data contribution in the Stackelberg equilibrium will satisfy the crowdsourcer’s $t\rho$ and $\overline{t\rho}$ bounds, i.e., $t\rho \leq t_{i} \rho_{i} \leq \overline{t\rho}$:

$$t\rho + \delta_{i} R_{l} \left( \frac{\left| Q_{l} \right| - 1}{\sum_{j \in Q_{l}} \frac{c_{j}}{\rho_{j}}} \right)^{2} \leq \overline{t\rho}, \tag{18}$$

where $\delta_{i} = \frac{\sum_{j \in Q_{l}} \frac{c_{j}}{\rho_{j}}}{\left| Q_{l} \right| - 1} \left( 1 - \frac{t\rho \sum_{j \in Q_{l}} \frac{c_{j}}{\rho_{j}}}{\left( \left| Q_{l} \right| - 1 \right) R_{l}} \right) - \frac{c_{i}}{\rho_{i}}$ and $\delta_{i} \geq 0$. 

15
Proof. We first show that (18) is a sufficient condition for \( t^*_i \rho_i \leq \bar{t}\rho \). From (18), we obtain:

\[
\left(\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}} R_{i1} - \frac{|Q_{i2}| - 1}{\sum_{j \in Q_{i2}} \frac{c_j}{\rho_j}} R_{i2}\right) + t\rho + \delta_i R_{i1} \left(\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}\right)^2 \leq \bar{t}\rho
\]

\[
\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}} R_{i1} - \left[\frac{1}{|Q_{i1}| - 1} - \left(\frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1}\right)^2 \frac{t\rho}{R_{i1}} - \delta_i \right] \left(\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}\right)^2 R_{i1} \leq \bar{t}\rho
\]

\[
\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}} R_{i1} - \left[\frac{1}{|Q_{i1}| - 1} - \left(\frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1}\right)^2 \frac{t\rho}{R_{i1}} - \delta_i \right] \left(\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}\right)^2 R_{i1} \leq \bar{t}\rho
\]

\[
(i) \quad \frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}} R_{i1} - \left[\frac{1}{|Q_{i1}| - 1} - \left(\frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1}\right)^2 \frac{t\rho}{R_{i1}} - \delta_i \right] \left(\frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}\right)^2 R_{i1} \leq \bar{t}\rho
\]

\[
(ii) \quad \sum_{j \in Q_{i1}} t_j \rho_j - \frac{c_i}{\rho_i} R_{i1} \left(\sum_{j \in Q_{i1}} t_j \rho_j\right)^2 \leq \bar{t}\rho
\]

\[
(iii) \quad t^*_i \rho_i \leq \bar{t}\rho,
\]

where (i) we substitute \( \frac{c_i}{\rho_i} = \frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1} - \delta_i \), (ii) we substitute \( \sum_{j \in Q_{i1}} t_j \rho_j = \frac{|Q_{i1}| - 1}{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}} R_{i1} \) from (26), and (iii) we substitute \( t^*_i \rho_i = \sum_{j \in Q_{i1}} t_j \rho_j - \frac{c_i}{\rho_i} R_{i1} \left(\sum_{j \in Q_{i1}} t_j \rho_j\right)^2 \) from (25).

After proving that (18) is a sufficient condition for \( t^*_i \rho_i \leq \bar{t}\rho \), we show that (18) also leads to \( t^*_i \rho_i \geq \bar{t}\rho \). From (18), we have \( \frac{c_i}{\rho_i} \leq \frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1} \left(1 - \frac{\sum_{j \in Q_{i1}} \frac{c_j}{\rho_j}}{|Q_{i1}| - 1} R_{i1}\right) \), which by Lemma 11 leads to \( t^*_i \rho_i \geq \bar{t}\rho \). Hence, the proof is complete.

Remarks: The feasible \( \bar{t}\rho \) and \( \bar{t}\rho \) for each worker \( i \) vary according to the its sensing cost \( c_i \) and location granularity \( \rho_i \), which affect \( \delta_i \) (the difference between the maximum allowable \( \frac{c_i}{\rho_i} \) value in the set of participating workers and the worker \( i \)'s \( \frac{c_i}{\rho_i} \) value). When \( \delta_i = 0 \), there is strict equality in (18) and \( \bar{t}\rho \) equals \( \bar{t}\rho \). However, as \( \delta_i \) increases, we have \( \bar{t}\rho < \bar{t}\rho \). For a fixed reward \( R_i \), the feasible range between \( \bar{t}\rho \) and \( \bar{t}\rho \) increases as the number of participating workers \( Q_{i1} \) increases.

4.2 Achieving Pareto Efficiency

We examine the Pareto efficiency (see Definition 15) of the Stackelberg equilibrium point in our proposed Stackelberg incentive model and study how to achieve efficiency. At the Pareto efficient equilibrium point, the crowdsourcer’s rewards are allocated in such a way that it is not possible to reallocate them to increase the utility of any individual (including the crowdsourcer itself) without making at least one individual’s utility decrease. In other words, it implies an efficient allocation of resources.
Definition 15 (Pareto Efficiency). A strategy profile \((R^P, t^P)\) is Pareto efficient if there exists no other strategy \((R, t)\) where \(R \geq 0, t \geq 0\) such that:

\[
U_{CS}(R, t) \geq U_{CS}(R^P, t^P), \\
u_i(t; t_{-i}, R) \geq u_i(t^P_{-i}; t^P_{-i}, R^P), \quad \forall i \in I,
\]

with at least one strict inequality.

Theorem 16. The proposed (unbounded) Stackelberg game has a unique Stackelberg equilibrium \((R^{SE}, t^{SE})\) that may not be Pareto efficient.

Proof: See Appendix 6.

By Theorem 16, the proposed (unbounded) Stackelberg game has a unique Stackelberg equilibrium \((R^{SE}, t^{SE})\) that may not be Pareto efficient. To achieve efficiency of the Stackelberg equilibrium, we first define a social welfare function \(w(R, t)\) to be the weighted sum of the crowdsourcer and the workers’ utilities:

\[
w(R, t) = \gamma_{CS} U_{CS}(R; t) + \sum_{i \in I} \gamma_i u_i(t; t_{-i}, R_i)
= \gamma_{CS} \sum_{i \in I} \alpha_i t_i^\beta + \sum_{i \in I} \gamma_i \left( \frac{t_i}{\rho_i} \sum_{j \in Q_i} t_j \rho_j R_i - c_i t_i \right),
\]

for some weights \(\gamma_{CS}, \gamma_1, \ldots > 0\).

It is well-known that any allocation which maximizes a social welfare function is also Pareto efficient. Thus, to achieve efficiency of the Stackelberg equilibrium, we can introduce the penalty function:

\[
\Psi(t) = \sum_{i \in I} \gamma_i \left( \frac{t_i}{\rho_i} \sum_{j \in Q_i} t_j \rho_j R_i - c_i t_i \right)
\]
to the crowdsourcer’s utility function, i.e.,

\[
U_{CS}(R; t) = \gamma_{CS} \sum_{i \in I} \alpha_i t_i^\beta + \Psi(t).
\]

One way to encourage the crowdsourcer to maximize \(w(R, t)\) would be to introduce a third party regulator that can offer tax rebates proportional to the weighted sum of the worker’s utilities \(\Psi(t)\), thus contributing to the crowdsourcer’s utility. Note that the penalty function \(\Psi(t)\) is a convex function of \(R\) and hence does not affect the convexity of the crowdsourcer’s Problem 1. Therefore, the unique Stackelberg equilibrium solution still exists.

5. Simulation Study

To evaluate the performance of the proposed Stackelberg incentive model in a real-world spatial data sensing application, we design a spatial data sensing case study application where we assume that the crowdsourcer wishes to perform a spatial estimation task. To achieve this, the crowdsourcer requires data from mobile smartphone users (workers) located in a range of different locations. The greater the spatial diversity obtained in the data, in general the easier it will be for the crowdsourcer to perform the spatial estimation task. In particular, we will design the case study to undertake a challenge of spatial prediction. The dataset collected by the crowdsourcer will be obtained by application of our spatial
privacy-preserving data sharing mechanism based on our proposed Stackelberg incentive model (Algorithm 2), which attempts to maximize spatial coverage.

We consider a real-world mobile crowd sensing problem such as the spatial monitoring and prediction of environmental temperature Chen et al. (2015); Mun et al. (2009). This involves the crowdsourcer incentivizing and paying the workers for the data they collected from their spatially placed sensors. Using the collected dataset, the crowdsourcer wishes to make a spatial estimation, achieved by conducting a spatial regression based on a Gaussian process Rasmussen and Williams (2005), estimated from the collected data. We then evaluate the spatial regression performance of the proposed Stackelberg incentive model against two baseline non-game-theoretic incentive schemes that seek to maximize the spatial coverage of the collected dataset.

5.1 Baseline Coverage Metrics

We consider the two baseline coverage maximizing schemes: (i) the location-based incentive mechanism proposed in Jaimes et al. (2012), which maximizes a geometric disk coverage model, and (ii) the work in Xiong et al. (2016), which maximize a $k$-depth coverage model. The two coverage models (see Fig. 3) are detailed as follows.

5.1.1 Geometric Disk Coverage Metric

The geometric disk coverage scheme was proposed in Jaimes et al. (2012) to measure the coverage $c(x_i)$ of a sensor data from a precise (uncloaked) location $x_i$:

$$c(x_i) = \begin{cases} 1 & \text{if } ||x_i - x_j||_2 \leq r, \\ 0 & \text{otherwise,} \end{cases}$$

where $|| \cdot ||_2$ denotes the Euclidean distance and $r$ is the sensed radius of the sensor data.

To optimize the disk coverage metric, the crowdsourcer greedily buys the minimum amount of data $t$ from each worker in regions where $c(x_i) = 0$, starting with the cheapest worker first. For a fair comparison with our proposed Stackelberg incentive model, we let $c(x_i) = 1$ when $x_i \in l$ where $l$ is a partitioned region.

5.1.2 $k$-depth Coverage Metric

The following $k$-depth coverage model (and its variants) was proposed in Xiong et al. (2016) to measure coverage of a set of $N$ sensor data $t_1, \ldots, t_N$ from a region $l$ where the coverage
Figure 4: Partitioned regions of the Intel lab, which contains 54 sensors nodes.

\[ c(t_1, \ldots, t_N) = \min(N, k) \] or equivalently:

\[ c(t_1, \ldots, t_N) = \begin{cases} 
N & \text{if } N \leq k, \\
\frac{N}{k} & \text{otherwise},
\end{cases} \]

where \( k \) is the depth parameter.

To optimize the \( k \)-depth coverage metric, the crowdsourcer greedily buys the minimum amount of data \( t \) from \( k \) workers in each region where \( c(t_i) = 0 \), starting with the cheapest worker first.

### 5.2 Simulation Setup

We use the temperature measurements from the Intel lab dataset Madden (2004 (accessed June 1, 2017), which contains the temperature, humidity, light, voltage, connectivity, and location information collected from 54 sensor nodes deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. We partitioned the lab’s spatial area into the eight regions \( l \) as shown in Fig. 4. The proposed Stackelberg incentive model and the baseline coverage schemes are then used to purchase a subset of the available temperature data. We took a one-hour interval (from 01:00–02:00, 28/2/2004) from the dataset and apply the Gaussian process regression technique Rasmussen and Williams (2005) (a supervised learning technique for regression) to evaluate how the two coverage metrics correspond to the actual amount of predictive uncertainty. The (Gaussian) radial basis function (RBF) kernel Rasmussen and Williams (2005) is used for our Gaussian process regression.

We briefly introduce the main idea behind the Gaussian process regression. Given a set of \( n \) input training location vector \( \mathbf{x} \) and observations \( \mathbf{y} \), we assume that the observed \( \mathbf{y} \) are generated by some latent function \( f \) plus an independent and identically distributed Gaussian noise with zero mean and variance \( \sigma_y^2 \). Suppose there are \( n_s \) test points where we are interested in obtaining the predicted observation values. Let \( k(\cdot, \cdot) \) be a covariance function and let \( K(X, X_s) \) denote the \( n \times n_s \) (kernel) matrix of the covariances evaluated at all pairs of training and test points, and similarly for \( K(X, X), K(X_s, X_s) \) and \( K(X_s, X) \). The predictive (posterior) mean \( \hat{f}_s(x_s) \) and variance \( V(x_s) \) for the Gaussian process regression for a new input test vector \( x_s \) are given by:

\[
\begin{align*}
\hat{f}_s(x_s) &= K(X_s, X)[K(X, X) + \sigma_y^2 I]^{-1}\mathbf{y}, \\
V(x_s) &= K(X_s, X_s) - K(X_s, X)[K(X, X) + \sigma_y^2 I]^{-1}K(X, X_s).
\end{align*}
\tag{21}
\]

There are a number of publicly available libraries that implement the Gaussian process regression and we chose to use Python’s scikit-learn machine learning library Pedregosa et al. (2011) in our simulations.
5.2.1 Test Scenarios

The following two test scenarios were examined.

Scenario (I): Spatial regression of the two cross intersections of interest (second and fourth column from the left side of Fig. 4) where no sensor data is available. We assume that each worker $i$’s sensing cost $c_i$ is inversely proportional to its distance from the nearest cross intersection of interest, i.e., workers have a higher sensing cost if they are located near the two cross intersections and their location granularities $\rho_i$ are inversely proportional to the number of workers in their respective regions, i.e., workers located in denser regions are less privacy-sensitive. We use the predictive variance $V(x_*)$ of the spatial Gaussian process regression in the two cross intersections (points of interest) as the metric for comparison. Intuitively, a lower predictive variance implies better predictive performance.

Scenario (II): Spatial regression of the entire spatial area shown in Fig. 4. We assume that the workers’ sensing costs $c_i$ are uniformly selected from $[0, 0.25, 0.5, 0.75, 1]$ and their location granularities $\rho_i$ are inversely proportional to the number of workers in their respective regions. We conducted spatial Gaussian process regression to obtain the predicted (mean) temperature $\bar{f}_s(x_*)$ of all the sensor locations. We use the mean squared error (MSE) values (computed by taking the difference between the predicted temperature and the actual temperature measurement from the dataset) as the metric for comparison. Intuitively, a lower MSE value implies better predictive performance.

5.2.2 Simulation Parameters

We set the sensing cost $c_i \in (0, 1]$, location granularity $\rho_i \in [1, 2]$, budget $R_{\text{budget}} = 10$, $R_{l}^{\text{min}} = 0$ for all $l \in L$, $R_{l}^{\text{max}} = \infty$ for all $l \in L$, minimum data $t = 1$, and maximum data $\bar{t} = 3$. We chose a budget constraint $R_{\text{budget}}$ to limit the number of purchased data, i.e., the scheme cannot afford to purchase all the available temperature data. In addition, the amount of budget spent by all three schemes are capped at the same amount. In the two baseline coverage schemes, we offer each worker their sensing cost for each unit of data. We let the crowdsourcer’s system parameter $\alpha_i = \rho_i$ and $\beta = 0.1$ in our proposed Stackelberg incentive model. We varied the $k$ value of the baseline $k$-depth coverage scheme.

5.3 Simulation Results and Discussion

We now discuss the simulation results for the test scenarios (I) and (II).

Scenario (I): We list the predictive variances of the two baseline coverage schemes and the proposed Stackelberg incentive model at the two cross intersections of interest in Table 2, and the corresponding coverage scores in Table 3.

From the results, we observe that the proposed Stackelberg incentive model has better predictive variances in the two cross intersections of interest compared to the baseline disk coverage scheme. The predictive variances from the $k$-depth coverage scheme and the proposed Stackelberg incentive model are the same for all values of $k > 1$. This is because the sensing costs of the workers are the lowest when the workers are located far away from the two cross intersections of interest and under the limited budget constraint, the $k$-depth coverage scheme and the proposed Stackelberg incentive model will selected the same set of participating workers. Note that the used disk coverage scheme is equivalent to the $k$-depth coverage scheme when $k = 1$. This is due to the usage of the partitioned regions instead of the original geometric disk coverage areas, which are not applicable to our scenario. To visualize the locations of the participating workers and the predictive variances, we plot the heat map of the predictive variances for the baseline coverage schemes and the proposed
Table 2: Predictive standard deviation values for Scenario (I).

| Scheme | Point 1 | Point 2 |
|--------|---------|---------|
| Disk   | 0.3435  | 0.3597  |
| k-depth| 0.2707  | 0.2620  |
| Proposed| 0.2707| 0.2620 |

Table 3: Baseline coverage scores for Scenario (I).

| Scheme     | Score | Disk | k-depth (k = 2) |
|------------|-------|------|-----------------|
| Disk       | 8     | 8    |                 |
| k-depth    | 8     | 16   |                 |
| Proposed   | 8     | 16   |                 |

Stackelberg incentive model in Fig. 5. The two cross labels in the figures represent the two points of interest where no sensor data is available. As the disk coverage scheme only selects one worker per region, it does not collect enough data points compared to the other two schemes.

Scenario (II): We list the MSE values of the two baseline coverage schemes and the proposed Stackelberg incentive model for the entire spatial area of interest in Table 4, and the corresponding coverage scores in Table 5. From the results, we observe that the proposed Stackelberg incentive model has better predictive performance (lower MSE value) in the entire spatial area of interest compared to the two baseline coverage schemes. While the baseline coverage schemes simply prioritize data from cheaper workers, our Stackelberg incentive model is able to offset the higher sensing costs of the workers with the $\alpha_i$ parameters. To visualize the location of the participating workers and the predicted mean temperature values, we plot the heat map of the predicted mean values for the baseline disc and k-depth schemes for $k = 3$ (where the MSE value is the lowest) and the proposed Stackelberg incentive model in Fig. 6.

Figure 5: Heatmap of the predictive standard deviation for (i) disk, and (ii) the k-depth and proposed Stackelberg incentive models - Scenario (I).
Table 4: Mean squared error values for Scenario (II).

| Scheme          | Values          | MSE  | Standard Deviation |
|-----------------|-----------------|------|--------------------|
| Disk            |                 | 35.0554 | 0.5228              |
| $k$-depth ($k = 2$) |                 | 26.5106 | 0.3338              |
| $k$-depth ($k = 3$) |                 | 26.3102 | 0.2832              |
| $k$-depth ($k = 4$) |                 | 32.9805 | 0.3034              |
| $k$-depth ($k = 5$) |                 | 39.4518 | 0.3431              |
| Proposed        |                 | 25.5666 | 0.2856              |

Table 5: Baseline coverage scores for Scenario (II).

| Scheme          | Score          | Disk | $k$-depth ($k = 2$) | $k$-depth ($k = 3$) | $k$-depth ($k = 4$) | $k$-depth ($k = 5$) |
|-----------------|----------------|------|---------------------|---------------------|---------------------|---------------------|
| Disk            |                 | 8    | 8                   | 8                   | 8                   | 8                   |
| $k$-depth       |                 | 8    | 16                  | 22                  | 23                  | 23                  |
| Proposed        |                 | 6    | 12                  | 18                  | 22                  | 23                  |

Figure 6: Heatmap of the predictive means in Scenario (II). Left: disk; Middle: $k$-depth; Right: proposed Stackelberg incentive model.
5.3.1 Accommodating Location Uncertainty

The Gaussian process regression technique accommodates location uncertainty of the workers’ sensing data due to the use of cloaking regions. Assuming that the input locations of the sensing data $x$ are corrupted by i.i.d. Gaussian noise with the noise variance set to the square of the approximated radius of the cloaking regions, the Gaussian process regression model proposed in Mchutchon and Rasmussen (2011) can be applied to account for location uncertainty. Mainly, an additional corrective term proportional to the gradient of the posterior mean can be added into the noise term in (21) to account for the location uncertainty of the training inputs.

6. Conclusion

We designed a privacy-aware Stackelberg incentive mechanism that improves the spatial coverage of the collected dataset. Our proposed incentive model is privacy-aware, in that it allows privacy-sensitive mobile smartphone users to submit coarse-grained (or quantized) location information to the crowdsourcer. We studied the properties of the proposed Stackelberg incentive model analytically and presented efficient algorithmic solutions. We also extended the basic model to accommodate bounds on the users’ data contributions and studied how Pareto efficiency can be achieved. We showed via simulations using a real-world sensing dataset that our proposed incentive model produced greater spatial diversity in sourced data, leading to better model predictive performance compared to two other coverage-maximizing schemes that maximize a different coverage metric. For future work, it would be interesting to extend our (static game) Stackelberg model for dynamic games played over a period of time where the smartphone users are allowed to move between regions. While our Stackelberg equilibrium is stable in the studied static model played by the crowdsourcer and the mobile smartphone users in one time period, a different notation of equilibrium needs to be considered for the dynamic setting.

References

Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In Proc. ACM Conference on Computer and Communications Security (CCS), pages 901–914, 2013.

P. Y. Chen, S. M. Cheng, P. S. Ting, C. W. Lien, and F. J. Chu. When crowdsourcing meets mobile sensing: a social network perspective. IEEE Commun. Mag., 53(10):157–163, Oct. 2015. ISSN 0163-6804. doi: 10.1109/MCOM.2015.7295478.

Reynold Cheng, Yu Zhang, Elisa Bertino, and Sunil Prabhakar. Preserving user location privacy in mobile data management infrastructures. In Proc. Int. Workshop on Privacy Enhancing Technologies (PET), pages 393–412, Jun. 2006.

Z. Feng, Y. Zhu, Q. Zhang, L. M. Ni, and A. V. Vasilakos. TRAC: truthful auction for location-aware collaborative sensing in mobile crowdsourcing. In Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM), pages 1231–1239, Apr. 2014. doi: 10.1109/INFOCOM.2014.6848055.
Bin Guo, Zhu Wang, Zhiwen Yu, Yu Wang, Neil Y. Yen, Ruihe Huang, and Xingshe Zhou. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Comput. Surv.*, 48(1):7–31, August 2015. ISSN 0360-0300.

L. G. Jaimes, I. Vergara-Laurens, and M. A. Labrador. A location-based incentive mechanism for participatory sensing systems with budget constraints. In *Proc. IEEE Int. Conf. Pervasive Comput. and Commun.*, pages 103–108, March 2012. doi: 10.1109/PerCom.2012.6199855.

H. Kido, Y. Yanagisawa, and T. Satoh. An anonymous communication technique using dummies for location-based services. In *Proc. Int. Conf. Pervasive Services.*, pages 88–97, Jul. 2005.

Jing Yang Koh, Gareth W. Peters, Derek Leong, Ido Nevat, and Wai-Choong Wong. Privacy-aware incentive mechanism for mobile crowd sensing. In *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017.

Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. Modeling users’ mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proc. Symp. Usable Privacy and Security (SOUPS)*, Jul. 2014.

Shuyun Luo, Yongmei Sun, Yufeng Ji, and Dong Zhao. Stackelberg game based incentive mechanisms for multiple collaborative tasks in mobile crowdsourcing. *Mobile Networks and Applications*, 21(3):506–522, 2016a. ISSN 1572-8153. doi: 10.1007/s11036-015-0659-3. URL http://dx.doi.org/10.1007/s11036-015-0659-3

Tie Luo, Sajal K. Das, Hwee Pink Tan, and Lirong Xia. Incentive mechanism design for crowdsourcing: An all-pay auction approach. *ACM Trans. Intell. Syst. Technol.*, 7(3):35:1–35:26, February 2016b. ISSN 2157-6904. doi: 10.1145/2837029. URL http://doi.acm.org/10.1145/2837029.

Samuel Madden. *Intel Lab Data*, 2004 (accessed June 1, 2017). [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html.

Andrew Mchutchon and Carl E. Rasmussen. Gaussian process training with input noise. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1341–1349. 2011.

Sarah E. Minson, Benjamin A. Brooks, Craig L. Glennie, Jessica R. Murray, John O. Langbein, Susan E. Owen, Thomas H. Heaton, Robert A. Iannucci, and Darren L. Hauser. Crowdsourced earthquake early warning. *Science Advances*, 1(3), Apr. 2015.

Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Péter Boda. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proc. ACM Mobile Syst., Applications, and Services (MobiSys)*, pages 55–68, 2009.

Daisuke Murakami, Gareth W Peters, Yoshiki Yamagata, and Tomoko Matsui. Participatory sensing data tweets for micro-urban real-time resiliency monitoring and risk management. *IEEE Access*, 4:347–372, 2016.

Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proc. ACM Conf. Electronic Commerce (EC)*, pages 774–789, 2012.
A. Overeemn, J. R. Robinson, H. Leijnse, G. J. Steeneveld, B. P. Horn, and R. Uijlenhoet. Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophysical Research Letters*, 40(15):4081–4085, Aug. 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Stephen G. Powell and Robert J. Batt. *Modeling for Insight: A Master Class for Business Analysts*. Hoboken, NJ: Wiley, 2008.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA: MIT Press, 2005. ISBN 026218253X.

Francesco Restuccia, Sajal K. Das, and Jamie Payton. Incentive mechanisms for participatory sensing: Survey and research challenges. *ACM Trans. Sen. Netw.*, 12(2):13:1–13:40, April 2016. ISSN 1550-4859. doi: 10.1145/2888398. URL http://doi.acm.org/10.1145/2888398.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965.

Nihar Shah and Dengyong Zhou. No oops, you won’t do it again: Mechanisms for self-correction in crowdsourcing. In *International Conference on Machine Learning*, pages 1–10, 2016.

Nihar Shah, Dengyong Zhou, and Yuval Peres. Approval voting and incentives in crowdsourcing. In *International Conference on Machine Learning*, pages 10–19, 2015.

Adish Singla and Andreas Krause. Truthful incentives for privacy tradeoff: Mechanisms for data gathering in community sensing. In *Proc. Int. Conf. Machine Learning (ICML)*, Jun. 2013.

Arvind Thiagarajan, Lenin Ravindrath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones. In *Proc. ACM Conf. Embedded Networked Sensor Syst. (SenSys)*, pages 85–98, 2009. ISBN 978-1-60558-519-2. doi: 10.1145/1644038.1644048. URL http://doi.acm.org/10.1145/1644038.1644048.

H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. E. Barnes. iCrowd: Near-optimal task allocation for piggyback crowdsensing. *IEEE Trans. Mobile Comput.*, 15(8): 2010–2022, Aug. 2016. ISSN 1536-1233. doi: 10.1109/TMC.2015.2483505.

D. Yang, X. Fang, and G. Xue. Truthful incentive mechanisms for k-anonymity location privacy. In *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, pages 2994–3002, 2013. doi: 10.1109/INFCOM.2013.6567111.

Dejun Yang, Guoliang Xue, Xi Fang, and Jian Tang. Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing. In *Proc. ACM Mobile Comput. and Networking (MobiCom)*, pages 173–184, 2012. doi: 10.1145/2348543.2348567. URL http://doi.acm.org/10.1145/2348543.2348567.
Algorithm 1: Compute the Nash equilibrium solution of the Followers game \((\mathcal{I}, \{t_i\}_{i\in \mathcal{I}}, \{u_i\}_{i\in \mathcal{I}})\).

1 function SolveFollowersGame(\(c, \rho, l, R\)):
   \textbf{Input} : sensing costs \(c_1,...,N\), location granularities \(\rho_1,...,N\), workers' regions \(l_1,...,N\), rewards \(R_1,...,L\).
   \textbf{Output}: data sold to crowdsourcer \(t^*_1,...,N\).
   \textbf{foreach} region \(l \in \mathcal{L}\) \textbf{do}
   3 Sort workers in region \(l\) according to their privacy-weighted cost \(\frac{c_i}{\rho_i}\) in ascending order where \(\frac{c_i}{\rho_i} \leq \frac{c_{i+1}}{\rho_{i+1}}\).
   4 Let \(Q_l = \{1, 2\}\) be the set of participating workers with \(t^*_i > 0\).
   5 Set \(Q_l \leftarrow Q_l \cup \{i\}\) for each worker \(i = 3, ..., \text{in region } l\) if the condition in (7) is met.
   (Note: the looping can stop at the \(i\)th step when the condition is not met.)
   6 Set \(t^*_i\) according to (6) for all workers in region \(l\).
   \textbf{end}

Algorithm 2: Compute the bounded Nash equilibrium solution for the Followers game \((\mathcal{I}, \{t_i\}_{i\in \mathcal{I}}, \{u_i\}_{i\in \mathcal{I}})\).

1 function SolveBoundedFollowersGame(\(c, \rho, l, R, t_\rho, \overline{t_\rho}\)):
   \textbf{Input} : sensing costs \(c_1,...,N\), location granularities \(\rho_1,...,N\), workers' regions \(l_1,...,N\), rewards \(R_1,...,L\).
   \textbf{Output}: data sold to crowdsourcer \(t^*_1,...,N\).
   \textbf{foreach} region \(l \in \mathcal{L}\) \textbf{do}
   3 Compute unbounded \(t_i\) for all workers in region \(l\) using Algorithm 1.
   4 if there exists more than one participating worker \(i\) with \(t_i \rho_i \geq t_\rho\) then
   5 Sort workers in region \(l\) according to their privacy-weighted cost \(\frac{c_i}{\rho_i}\) in ascending order where \(\frac{c_i}{\rho_i} \leq \frac{c_{i+1}}{\rho_{i+1}}\).
   6 Let \(Q_l = \{1, 2\}\) be the set of participating workers with \(t^*_i > 0\).
   7 Set \(Q_l \leftarrow Q_l \cup \{i\}\) for each worker \(i = 3, ..., \text{in region } l\) if condition (13) of Theorem 10 is met.
   (Note: the looping can stop at the \(i\)th step when the condition is not met.)
   8 Set \(t^*_i\) according to (6) for all workers in region \(l\).
   9 else
   10 Set \(t^*_i = 0\) for all workers \(i\) in region \(l\).
   11 end
.1 Proof of Lemma 2

A unique Nash equilibrium exists in the Followers game if for all \( i \in \mathcal{I} \) Rosen (1965): (i) the domain of the workers’ strategy set \( \{ t_i \}_{i \in \mathcal{I}} \) is convex and compact, and (ii) \( u_i \) is continuous and strictly concave in \( t_i \). Indeed, the domain of the workers’ strategy set \( \{ t_i \}_{i \in \mathcal{I}} \) is convex and compact since \( t_i \) is assumed to be bounded, and \( u_i \) is continuous in \( t_i \), and strictly concave in \( t_i \) as \( \frac{\partial^2 u_i}{\partial t_i^2} < 0 \):

\[
\frac{\partial u_i}{\partial t_i} = \frac{\rho_i}{\left( \sum_{j \in \mathcal{Q}_i, j \neq i} t_j \rho_j \right)^2} R_i - c_i,
\]

\[
\frac{\partial^2 u_i}{\partial t_i^2} = -2\rho_i^2 \sum_{j \in \mathcal{Q}_i, j \neq i} t_j \rho_j \left( \sum_{j \in \mathcal{Q}_i} t_j \rho_j \right)^3 R_i < 0. \tag{22}
\]

Therefore, a Nash equilibrium exists in the Followers game. \( \square \)

.2 Proof of Theorem 3

By Lemma 2, there exists a unique strategy profile that maximizes the utility of each worker given the strategies of the other workers. Thus, if each worker \( i \) plays its best response strategy, it will achieve the unique Nash equilibrium point \( t_i^* \). To prove Theorem 3, we derive \( t_i^* \) by solving \( \frac{\partial u_i}{\partial t_i} = 0 \) to obtain:

\[
\rho_i R_i \sum_{j \in \mathcal{Q}_i, j \neq i} t_j^* \rho_j = c_i \left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \right)^2. \tag{23}
\]

Next, we manipulate the expression in (23) to obtain:

\[
\left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j - t_i^* \rho_i \right) \rho_i R_i = c_i \left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \right)^2 \tag{24}
\]

\[
\Rightarrow t_i^* \rho_i = \left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \right) \left[ 1 - \frac{c_i}{\rho_i R_i} \left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \right) \right] \tag{25}
\]

\[
\Rightarrow \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j = |\mathcal{Q}_i| \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j - \frac{1}{R_i} \left( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \right)^2 \sum_{j \in \mathcal{Q}_i} \frac{c_j}{\rho_j} \tag{26}
\]

where (i) we express \( t_i^* \rho_i \) in terms of the other \( t_j^* \rho_j \) values, (ii) we sum up the \( t_i^* \rho_i \) values in (25) for all participating workers in region \( l_i \), and (iii) we divide the previous expression by \( \sum_{j \in \mathcal{Q}_i} t_j^* \rho_j \).

Finally, we substitute (26) into (25) to obtain the unique Nash equilibrium point for each worker \( i \) as required. \( \square \)
.3 Proof of Lemma 4

To prove Lemma 4, we show that the set of participating workers $Q_l$ computed by Algorithm 1 always satisfies (7) for all regions $l \in L$.

Consider the region $l$. Suppose the constraint in (7) is not satisfied, then the following constraints must be true:

$$\frac{c_i}{\rho_i} \geq \sum_{j \in Q_l; j \neq i} \frac{c_j}{\rho_j} + \frac{c_i}{\rho_i} \left| Q_l \right| - 1,$$

$$\frac{c_i}{\rho_i} \geq \sum_{j \in Q_l; j \neq i} \frac{c_j}{\rho_j} \left| Q_l \right| - 2. \tag{27}$$

From (27), it is obvious that if $\frac{c_k}{\rho_k} \geq \frac{c_i}{\rho_i}$ and $\frac{c_i}{\rho_i} \geq \sum_{j \in Q_l; j \neq i} \frac{c_j}{\rho_j} \left| Q_l \right| - 2$, then $\frac{c_k}{\rho_k} \geq \sum_{j \in Q_l; j \neq i} \frac{c_j}{\rho_j} \left| Q_l \right| - 2$. Hence, the greedy computation of the set of participating workers $Q_l$ from the sorted $\frac{c_i}{\rho_i}$ values in Algorithm 1 does not affect the set of participating workers $Q_l$ that achieves the unique Nash equilibrium solution of the Followers game. $\square$

.4 Proof of Theorem 5

To show that the obtained solution from Algorithm 1 is the unique Nash equilibrium (NE) solution of the workers, we prove that the $\frac{\partial u_i}{\partial t_i} = 0$ (stationary point) condition given in (24) is satisfied by the set of participating workers $i \in Q$. By Lemma 2, there exists a unique NE in the followers game. Hence, any stationary point is the unique NE point for the workers. From Algorithm 1, we have $t_i^* = \frac{|Q_l| - 1}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} \left( 1 - \frac{|Q_l| - 1}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} \right) \frac{R_i}{\rho_i}$ if $i \in Q_l$, and $t_i^* = 0$ otherwise. In addition, we have the expression $\sum_{j \in Q_l} t_j^* \rho_j = \frac{|Q_l| - 1}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} R_i$ from (26). We substitute the expressions for $t_i^*$ and $\sum_{j \in Q_l} t_j^* \rho_j$ into the $\frac{\partial u_i}{\partial t_i}$ expression in (24) to obtain the following equality:

$$\frac{\left( |Q_l| - 1 \right) R_i}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} \left( \frac{|Q_l| - 1}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} \right) R_i = \frac{c_i}{\rho_i} \left[ \frac{|Q_l| - 1}{\sum_{j \in Q_l} \frac{c_j}{\rho_j}} R_i \right]^2. \tag{28}$$

Since (28) satisfies the expression for $\frac{\partial u_i}{\partial t_i} = 0$, we conclude that Algorithm 1 correctly outputs the unique Nash equilibrium solution of the workers in $Q_l$. Also, by Lemma 4, Algorithm 1 also correctly computes the set of participating workers $Q_l$ as used in (28). $\square$
.5 Proof of Lemma 8

From (11), we have the following expression for all workers \( i \in Q_i \):

\[
\frac{\partial u_i}{\partial c_i} = -(|Q_i| - 1) \frac{\left( \sum_{j \in Q_i, j \neq i} c_j \rho_j \right)^2 + \frac{c_i}{\rho_i} \left( |Q_i| - 2 \right) \left( \frac{c_i}{\rho_i} \right) + \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} - 2|Q_i| \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j}}{\left( \sum_{j \in Q_i} \frac{c_j}{\rho_j} \right)^3} R_i,
\]

\[
\frac{\partial^2 u_i}{\partial (c_i)^2} = 2(|Q_i| - 1) \left( \sum_{j \in Q_i, j \neq i} c_j \rho_j \right)^2 + \frac{c_i}{\rho_i} \left( |Q_i| - 2 \right) \left( \frac{c_i}{\rho_i} \right) + \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} - (2|Q_i| - 1) \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} R_i.
\]

From (29), we set \( \frac{\partial u_i}{\partial c_i} = 0 \) to obtain the critical point:

\[
|Q_i| \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} = \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right)^2 + \frac{c_i}{\rho_i} \left( |Q_i| - 2 \right) \left( \frac{c_i}{\rho_i} \right) + \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right)^2.
\]

\[
\frac{c_i}{\rho_i} = \frac{|Q_i| \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} - \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right)^2}{(2|Q_i| - 1) \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j}}.
\]

We substitute the expression of the critical point derived in (30) into the \( \frac{\partial^2 u_i}{\partial (c_i)^2} \) expression in (29) to obtain:

\[
\frac{\partial^2 u_i}{\partial (c_i)^2} \propto \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right)^2 + |Q_i| \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} - \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right)^2 - (2|Q_i| - 1) \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \left( \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \right).
\]

\[
\propto \left( \frac{c_i}{\rho_i} \right) \sum_{j \in Q_i, j \neq i} \frac{c_j}{\rho_j} \left( |Q_i| - (2|Q_i| - 1) \right),
\]

\[
\propto 1 - |Q_i|.
\]

The denominator term \( \sum_{j \in Q_i} \frac{c_j}{\rho_j} \) in the \( \frac{\partial^2 u_i}{\partial (c_i)^2} \) expression from (29) is positive due to both \( c_i \) and \( \rho_i \) being positive. Since we assume that \( |Q_i| > 2 \), then we have \( \frac{\partial^2 u_i}{\partial (c_i)^2} < 0 \). This leads us to conclude that the critical point is a maximum point. Hence, the proof is complete. □

.6 Proof of Theorem 16

To prove Theorem 16, we make the following two claims.

Claim 1. Suppose that the \( \alpha_i, \rho_i, \) and \( c_i \) values are constant for all workers \( i \in I \), the Stackelberg equilibrium of the proposed Stackelberg incentive model is Pareto efficient.
Proof. To prove the claim, we use the following key observations:

\[
\sum_{j \in Q_i} u_j(t_j; t_{-j}, R_i) = R_i - \sum_{j \in Q_i} c t_j.
\]  
(32)

\[
U_{CS}(R; t) = \alpha \sum_{i \in I} t_i^\beta.
\]  
(33)

Consider a strategy profile \((R, t) \neq (R^{SE}, t^{SE})\). If \(U_{CS}(R; t) > U_{CS}(R^{SE}; t^{SE})\), then \(\exists i\) where

\[
u_i(t_i; t_{-i}, R_i) < \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE})\].

It can be shown that the previous statement is true since

\[
\sum_{j \in Q_i} t_j > \sum_{j \in Q_i} t_j^{SE}
\]

is a necessary condition for \(U_{CS}(R; t) > U_{CS}(R^{SE}; t^{SE})\) when the \(\alpha_i, \rho_i, \) and \(c_i\) values are constant for all workers \(i \in I\). From (32), we know that \(\sum_{j \in Q_i} u_j(t_j; t_{-j}, R_i)\)

is inversely proportional to \(\sum_{j \in Q_i} c t_j\). This means that if \(\sum_{j \in Q_i} t_j > \sum_{j \in Q_i} t_j^{SE}\) is true, then

we have \(\sum_{j \in Q_i} u_j(t_j; t_{-j}, R_i) < \sum_{j \in Q_i} u_j(t_j^{SE}; t_{-j}^{SE}, R_i^{SE})\). Hence, we conclude that \(\exists i\) where

\[
u_i(t_i; t_{-i}, R_i) < \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE}).\]

Similarly, it can be shown that if \(\exists i\) where \(\nu_i(t_i; t_{-i}, R_i) > \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE})\) and \(\nu_i(t_i; t_{-i}, R_i) \geq \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE})\), \(\forall i \in I\), then \(U_{CS}(R; t) < U_{CS}(R^{SE}; t^{SE})\). This is because if \(\exists i\) where \(\nu_i(t_i; t_{-i}, R_i) > \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE})\) and \(\nu_i(t_i; t_{-i}, R_i) \geq \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE})\), \(\forall i \in I\), then we have

\[
\sum_{j \in Q_i} u_j(t_j; t_{-j}, R_i) > \sum_{j \in Q_i} u_j(t_j^{SE}; t_{-j}^{SE}, R_i^{SE}).
\]

This is only possible if \(\sum_{j \in Q_i} c t_j < \sum_{j \in Q_i} c t_j^{SE}\), which means \(\sum_{j \in Q_i} t_j < \sum_{j \in Q_i} t_j^{SE}\). Therefore, from (33), we conclude that

\[
U_{CS}(R; t) < U_{CS}(R^{SE}; t^{SE}).
\]

Hence, the proof is complete.

\[\blacksquare\]

Claim 2. The proposed Stackelberg incentive model may not have a Pareto efficient Stackelberg equilibrium. In other words, suppose that \((R, t) \neq (R^{SE}, t^{SE})\) and \(U_{CS}(R; t) > U_{CS}(R^{SE}; t^{SE})\), we have

\[
u_i(t_i; t_{-i}, R_i) \geq \nu_i(t_i^{SE}; t_{-i}^{SE}, R_i^{SE}), \forall i \in I.
\]

Proof. Consider the scenario where there are 2 regions \(l_1\) and \(l_3\) with 2 workers each. Let \(\rho_1 = 1, c_1 = 1\) for the workers in the 2 regions. Let \(l_1 = l_2\) and \(l_3 = l_4\), \(\alpha_1 = \alpha_2, \alpha_3 = \alpha_4\), and \(\alpha_1 < \alpha_3\). Intuitively, given that the worker costs are the same but \(\alpha_1 < \alpha_3\), then \(R_i^{SE} < R_i^{SE}\).

Suppose we have \(R_3 > R_3^{SE}\) and \(R_1 < R_1^{SE}\) (recall that the total rewards is bounded, so \(R_i\) must decrease if \(R_3\) increases). Let \(R_3 = R_3^{SE} + \Delta\) and \(R_1 = R_1^{SE} - \Delta\) where \(\Delta > 0\).

First, we obtain the closed-form expression for \(t_i\) at the Stackelberg equilibrium. We substitute the \(\rho_i\) and \(c_i\) values into (6) and obtain:

\[
t_i^{SE} = \frac{R_i^{SE}}{2} (1 - \frac{1}{2}) = \frac{R_i^{SE}}{4}, \forall i \in I.
\]  
(34)
Next, we substitute the $t_i$ expression in (34) into (3) to obtain

$$u_i(t_i^{SE}; t_{-i}^{SE}, R_{li}^{SE}) = \frac{R_{li}^{SE}}{2} - \frac{R_{li}^{SE}}{4} = \frac{R_{li}^{SE}}{4}. \quad (35)$$

Suppose $u_1(t_i; t_{-i}, R_{li}) \geq u_1(t_i^{SE}; t_{-i}^{SE}, R_{li}^{SE})$, from (3), we have:

$$t_1 = \frac{R_{li}^{SE} - 2\Delta}{4} - \Delta', \quad (36)$$

where $\Delta' \geq 0$.

Similarly, if $u_3(t_i; t_{-i}, R_{li}) \geq u_3(t_i^{SE}; t_{-i}^{SE}, R_{li}^{SE})$, from (3), we have:

$$t_3 = \frac{R_{li}^{SE} + 2\Delta}{4} - \Delta', \quad (37)$$

where $\Delta' \geq 0$.

If $U_{CS}(R; t) > U_{CS}(R^{SE}; t^{SE})$, and $u_i(t_i; t_{-i}, R_{li}) \geq u_i(t_i^{SE}; t_{-i}^{SE}, R_{li}^{SE}), \forall i \in I$ is true, then the following condition on $U_{CS}$ must be true. From (8), we have:

$$\alpha_1 t_1^\beta + \alpha_2 t_2^\beta + \alpha_3 t_3^\beta + \alpha_4 t_4^\beta > \alpha_1 (t_1^{SE})^\beta + \alpha_2 (t_2^{SE})^\beta + \alpha_3 (t_3^{SE})^\beta + \alpha_4 (t_4^{SE})^\beta,$$

$$\alpha_1 \left(\frac{R_{li}^{SE} - 2\Delta}{4} - \Delta'\right)^\beta + \alpha_3 \left(\frac{R_{li}^{SE} + 2\Delta}{4} - \Delta'\right)^\beta > \alpha_1 \left(\frac{R_{li}^{SE}}{4}\right)^\beta + \alpha_3 \left(\frac{R_{li}^{SE}}{4}\right)^\beta. \quad (38)$$

Therefore, Claim 2 shows a scenario where the proposed Stackelberg equilibrium may not be Pareto efficient for the crowdsourcer.