On the Data Quality in Privacy-Preserving Mobile Crowdsensing Systems with Untruthful Reporting

Cong Zhao, Shusen Yang, Senior Member, IEEE, and Julie A. McCann

Abstract—The proliferation of mobile smart devices with ever improving sensing capacities means that human-centric Mobile Crowdsensing Systems (MCSs) can economically provide a large scale and flexible sensing solution. The use of personal mobile devices is a sensitive issue, therefore it is mandatory for practical MCSs to preserve private information (the user’s true identity, precise location, etc.) while collecting the required sensing data. However, well intentioned privacy protection techniques also conceal autonomous, or even malicious, behaviors of device owners (termed as self-interested), where the objectivity and accuracy of crowdsensing data can therefore be severely threatened. The issue of data quality due to untruthful reporting in privacy-preserving MCSs has been yet to produce solutions. Bringing together game theory, algorithmic mechanism design, and truth discovery, we develop a mechanism to guarantee and enhance the quality of crowdsensing data without jeopardizing the privacy of MCS participants. Together with solid theoretical justifications, we evaluate the performance of our proposal with extensive real-world MCS trace-driven simulations. Experimental results demonstrate the effectiveness of our mechanism on both enhancing the quality of the crowdsensing data and eliminating the motivation of MCS participants, even when their privacy is well protected, to report untruthfully.

Index Terms—Mobile Crowdsensing Systems, Privacy Preservation, Data Quality, Untruthful Reporting.

1 INTRODUCTION

Technological developments and increased uptake of personal mobile smart devices (e.g. smart phones, wearables, and vehicles) is spurring the emergence of various human-centric Mobile Crowdsensing Systems (MCSs) [1]. Here people contribute information through the use of mobile devices equipped with enriched on-board sensing components (e.g. camera, microphone, compass, and accelerometer), and are becoming more widely used in fields such as urban traffic and road monitoring [2], [3], environment monitoring [4], [5], mobile advertising [6], [7], public healthcare [8], [9], etc. Leveraging such pervasive mobile devices, while utilising their enhanced spatiotemporal coverage, the MCS has become key to flexible economical large-scale sensing [10]. However, one of the essential properties of MCSs is that the performance of crowdsensing services highly depends on the less-predictive behaviors of the people involved in the sensing activity (i.e. the temporally recruited human participants). The introduction of large scale human interactions raises the conflicting issues of both personal private information protection [11] and crowdsensing data quality guarantees [12] that have to be addressed in practical MCS applications.

Specifically, one of the essential requirements for the Crowdsensing Service Provider (CSP) is to guarantee that the private information of all MCS participants (i.e. mobile device owners who participate in sensing tasks) is properly protected [11]. In fact, considerable efforts are being made to preserve participant’s privacy in terms of their identity [13], [14], precise location [15], [16], [17], and plain-text sensor content [18], [19], [20], etc. However, with the adoption of such privacy-preserving mechanisms individual behaviors of the so called self-interested participants (or indeed their strategies) are inevitably ‘blurred’ from the CSP’s point of view, and this potentially challenges the accuracy of the crowdsensing service. For a specific sensing task, due to the lack of accountable historical information (to protect participant’s privacy), it is difficult for the CSP to identify the reliable participants who are willing to provide objective and accurate data (also referred to as ‘high-quality data’ in this paper) among all of the indistinguishable candidates, and it is non-trivial for the CSP to evaluate the quality of received data when there exists neither referable ground truth nor a proof of data truthfulness (indicating that the participants have not tampered with their reported data). Obviously, the controversy between privacy preservation and data quality has to be explicitly addressed to extract high-quality data from increasingly ‘cloaked’ participants, which is challenging but critical to MCSs [21].

As an active research topic, the issue of data quality [12] has been attracting significant interests of researchers in the MCS field, where the understanding of the quality of crowdsensing data is multi-fold. The majority of existing research [22], [23], [24], [25], [26], [27] considers the quality...
of received data as an intrinsic and relatively stable property determined by objective factors such as the reporting participant’s geological location, battery level, transmission bandwidth, etc. However, such a notion (similar to Quality of Service in traditional sensing systems) cannot reflect the impact of participant’s subjective behaviors (particularly malicious behaviors [12], [28]) on the quality of the crowdsensed data. As far as we are aware, limited research results [29], [30], [31] discuss quality-driven crowdsensing that protects against untruthful sensing data reporting. In [29], the CSP is able to estimate data quality according to the historical behavior profiles of all participants, which, unfortunately, is not achievable within the aforementioned privacy-preserving MCSs since there exists no accountable information about the behaviors of the cloaked participants. In [30], similarly, the approach to specifically recruit workers with more reliable data and the enforcement of truthful reporting would also be based on the accountable information about historical behaviors of the workers. In [31] a quality-driven incentivization mechanism is proposed for privacy-preserving MCSs, however the heuristic sensing duty allocation without budget limitations is actually not applicable or desirable for practical scenarios.

Considering the potential impact of a participant’s untruthful behavior regarding the quality of crowdsensing data in privacy-preserving MCSs, in our paper, we develop an incentive-based mechanism that forces self-interested crowdsensing participants to truthfully report accurate and reliable sensing observations to the CSP; who has no access to participant’s private information including true identities and precise location. Specifically, our mechanism contains three components: the quality-driven participant recruitment and optimized sensing duty allocation at the CSP; the mechanism design against untruthful reporting from self-interested participants; and the adaptive data quality estimation by participants themselves. The contributions of the proposed mechanism are summarized as follows:

1) As far as we know, our mechanism is the first work that systematically addresses the impact of untruthful reporting from well cloaked participants on the objectivity and accuracy of sensing data in privacy-preserving MCSs. In particular, our mechanism eliminates the motivation of rational participants to strategically tamper with the information reported to the CSP (including the self-estimated data quality indications as the application for participation, as well as the sensing data uploaded) to gain unfair advantages, yet without requiring any accountable information about participant’s historical behavior;

2) Beyond existing research, our introduction of the active sensing bias calibration by the participants aims at enhancing crowdsensing data quality from the source. Specifically, to get higher participating rewards with higher data quality, the participants gradually understand and calibrate their private and intrinsic ‘sensing bias’ based on the feedback from the CSP after each task, and then actively adjust their policy to upgrade the objectivity and accuracy of the information they report in future tasks;

3) Besides explicit theoretical analysis, we conduct extensive real-world trace-driven simulations to evaluate the performance of our mechanism. The results indicate that our mechanism manages to effectively enhance the quality of crowdsensing data (e.g. 56% reduction of the standard deviation of reported sensing data) and eliminate the motivation of cloaked participants to report untruthfully (e.g. 22.2% to 85.2% reduction of average sensing profit when there exist untruthful behaviors).

The remainder of this paper is organized as follows. We discuss the related work in Section 2. The system model and problem formulation are presented in Section 3. Section 4 provides an overview of our entire solution. In Section 5, we address the issue of quality-driven agent recruitment and sensing duty allocation. Section 6 specifically focuses on the mechanism design against untruthful reporting. The method of adaptive data quality self-estimation is presented in Section 7. Section 8 explicitly demonstrates our evaluation methodology, results, and discussions. Finally, we conclude this paper in Section 9.

2 RELATED WORK

In this section, we discuss the existing research efforts related to the data quality issue in privacy-preserving MCSs.

2.1 Proliferating MCS Applications

Flexible and economical large-scale sensing means that different human-centric MCSs are now being applied to various fields including (but not limited to) urban traffic and road monitoring [2], environment monitoring [5], mobile advertising [7], and public healthcare [9]. For urban traffic and road monitoring, Hu et al. propose a road sensing system SmartRoad in [2] for traffic regulator detection and identification based on participatory sensing data collected from GPS sensors from in-vehicle smartphones. As for the monitoring of environmental events, Rana et al. propose an end-to-end participatory urban noise mapping system Ear-Phone in [5], which addresses the issue of incomplete and random crowdsensed data with compressive sensing. Focusing on the field of mobile advertising, Wang et al. propose a method for audience-targeted billboard advertising by leveraging crowdsensing vehicle trajectory data in [7]. Considering public healthcare, Pryss et al. develop a mobile crowdsensing platform TrackYourTinnitus for the research and treatment of tinnitus in [9]. As shown, human-centric sensing is has the potential to aid daily life from different perspectives. In the meantime, however, issues include potential personal privacy leakage, undetermined quality of the crowdsensed data, and unreliable behavior of self-interested participants [12], which have to be explicitly addressed.

2.2 Privacy Preservation of MCS Participants

Realizing the inevitability of personal privacy preservation in the implementation of practical MCSs, a large amount of effort has been devoted to protect the accountable information of participants in terms of true identity [13], [14], precise geological location [15], [16], [17], and plain-text sensing
2.3 The Quality of Crowdsensing Data

Preservation has become a mandatory requirement for the prevention of potential inference attacks. In fact, personal privacy from each user is guaranteed as differentially private to mechanisms for MCSs, where the bidding of information is encrypted using a homomorphic cryptosystem. Yang etc. [20] propose an incentivization scheme with location-privacy preserving for MCSs, where k-anonymity is utilized to reduce the risk of location-privacy disclosure while the incurred information loss is mitigated.

Considering the potential privacy leakage from the sensing data reported by the participants, in [18], Miao etc. propose a novel privacy-preserving truth discovery framework PPTD that discovers the truth of crowdsensing tasks upon sensing data encrypted using a homomorphic cryptosystem. Yang etc. [19] propose a different private crowdsensing data aggregation scheme that protects the plain-text data by adding noise with designed distributions. In [20], Lin etc. construct a framework of privacy-preserving auction-based incentive mechanisms for MCSs, where the bidding of information from each user is guaranteed as differentially private to prevent potential inference attacks. In fact, personal privacy preservation has become a mandatory requirement for the implementation of practical human-centric MCSs in future.

2.4 Data Quality against Untruthful Reporting in MCSs

Currently, only limited research results [29], [30], [31] have discussed the data quality issue in MCSs where participants have the potential to deliberately report untruthful sensing observations or other information. In [29], Yang etc. propose a quality-based truth estimation and surplus sharing method for crowdsensing, where the data quality of each participant is quantified base on historical quality records using an unsupervised learning method. Unfortunately, this method is not applicable in privacy-preserving MCSs since no accountable information about the behaviors of participants is accessible to the service provider. In [30], Jin etc. develop an incentive mechanism INCEPTION for privacy-aware data aggregation in MCSs, which enables the CSP to recruit workers with more reliable data while enforcing truthful reporting from workers for the optimal profit. Unfortunately, this method cannot be directly applied when there exists no accountable information about worker’s historical behavior. In [31], Yang etc. develop a method to prevent dishonest reporting in privacy-preserving MCSs based on algorithmic mechanism design. However, none of the issues including budget-limited sensing duty allocation, incentive compatibility among multiple participants, or sensing bias calibration is considered, which strictly limits the method’s application in practical MCSs.

3 System Model and Problem Formulation

In this section, we present a general model of privacy-preserving MCSs, and explicitly describe our research problem. For clarity, frequently used notations in the paper are summarized in Table 1.

3.1 System Architecture

As illustrated in Fig.1, a typical MCS consists of a service provider s and a set of registered mobile agents \( A (|A| \geq 2) \). All agents communicate with s through different kinds of wireless technologies (e.g. WiFi and cellular). To preserve personal privacy, we assume that the true identities and precise geolocial locations of all agents are well protected by using the anonymous communication [13], [14] and spatial cloaking [15], [16] techniques, respectively. Each agent in \( A \) possesses a unique and dynamic pseudo-identity \( i \in N = \{1, 2, ..., N\} \) (where \( N = |A| \)) to communicate with s without revealing the true identity. Sensing data collected by each agent contain no precise geographical tag (i.e.
GPS coordinates) but only general regional information with variable spatial cloaking granularity, therefore reporting this sensed data will not reveal the agent’s precise movement pattern. In each MCS task \( \tau \), the responsibility of \( s \) is to accurately calculate the sensing truth \( \sigma_\tau \) based on all sensing observations reported by temporally recruited agents. All agents in \( N \) rationally participate in task \( \tau \) to get rewards from \( s \). If employed by \( s \), agent \( i \) completes the sensing obligation as requested, then reports an observation \( \hat{o}_i \) to \( s \), and is finally rewarded depending on the quality \( q_i \) of \( \hat{o}_i \).

Specifically, the general workflow of task \( \tau \) is as follows: \( s \) announces the sensing requirements of \( \tau \) (i.e. the type of sensing data required, the desired sensing location of that data \( L_\tau = (\text{lat}_\tau, \text{long}_\tau) \), and the effective sensing radius \( r_\tau \)) to all agents. Assuming that all agents in \( M = \{1, 2, ..., M\} \subseteq N \) (and \( M = |M| \)) want to participate in \( \tau \), each agent \( i \in M \) replies \( s \) with a self-estimated data quality indication \( \hat{q}_i \) (\( 0 < \hat{q}_i \leq 1 \)) as \( i \)'s participating application. With \( \hat{Q} = \{\hat{q}_1, \hat{q}_2, ..., \hat{q}_M\} \), \( s \) decides a sensing duty allocation \( D = \{d_1, d_2, ..., d_M\} \) considering \( \tau \)'s total sensing budget \( B \) > 0, where \( d_i \geq 0 \) (\( 1 \leq i \leq M \)) denotes the sensing duty allocated to agent \( i \). All employed (i.e. \( d_i > 0 \)) agents (or employees) in \( E = \{1, 2, ..., E\} \subseteq M \) (and \( E = |E| \)) then report their sensing observations \( \hat{O} = \{\hat{o}_1, \hat{o}_2, ..., \hat{o}_E\} \) according to \( D \). Based on \( \hat{O} \), \( s \) calculates the sensing truth \( o_\tau \) and evaluates the actual data quality of all employees, \( \hat{Q} = \{q_1, q_2, ..., q_E\} \), where \( 0 < q_i \leq 1 \) for \( \forall i \in E \). Besides, the relative contribution \( \delta_i \) is evaluated by \( s \) for each \( i \in E \) that reflects the difference between \( i \)'s expected (determined by \( \hat{Q} \)) and actual (determined by \( \hat{Q} \)) contributions. Finally, \( s \) rewards each agent \( i \in E \) with a sensing profit \( p_i \) (monetary credits) discriminatively considering \( D \) and \( \hat{Q} \), and replies \( \delta_i \) and \( o_\tau \) to \( i \). Based on \( \delta_i \) and \( o_\tau \) from \( s \), each agent \( i \in E \) calibrates the sensing bias \( b_i \) and adjusts \( \hat{q}_i \).

### 3.2 Problem Definition

In general, the essential goal of the crowdsensing service provider \( s \) is to accurately calculate the sensing truth \( o_\tau \) based on sensing observations \( \hat{O} = \{\hat{o}_1, \hat{o}_2, ..., \hat{o}_E\} \) from all employees in \( E \). Because of our privacy-preserving setting, \( s \) has no accountable information about each agent’s historical behaviors, therefore it is difficult for \( s \) to determine the quality of each received observation as well as the accuracy of the calculated sensing truth.

Considering the complexity of factors that will impact the quality of crowdsensing data, we intuitively divide our primary problem of data quality guarantee in privacy-preserving MCSs into the following three sub-problems under different assumptions with increasing generalities:

1) Firstly, we assume that each agent \( i \in M \) knows the actual data quality indication \( q_i \), and honestly applies to participate in \( \tau \) with \( q_i = \hat{q}_i \). In this case, the problem needs to be solved to guarantee the data quality is: how to recruit as many agents who can provide high quality data as possible with the limited total sensing budget, and how to maximize the participating profit of each employee for effective incentivization?

2) Then, with the established incentive method, we assume that each agent \( i \in M \) knows the actual data quality indication \( q_i \), but may apply to participate in \( \tau \) with \( q_i \neq \hat{q}_i \) for a potentially higher profit. In this case, the problem needs to be solved to guarantee the data quality is: how to enforce all agents truthfully reporting their actual data quality indications (and the following sensing observations)?

### Table 1: Notations.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>all mobile agents with pseudo-identities</td>
<td>( s )</td>
<td>the crowdsensing service provider</td>
</tr>
<tr>
<td>( \tau )</td>
<td>a specific crowdsensing task</td>
<td>( L_\tau )</td>
<td>the desired sensing location of ( \tau )</td>
</tr>
<tr>
<td>( r_\tau )</td>
<td>the effective sensing radius of ( \tau )</td>
<td>( B )</td>
<td>the total sensing budget of ( \tau )</td>
</tr>
<tr>
<td>( M )</td>
<td>all participating applicants of ( \tau )</td>
<td>( \hat{Q} )</td>
<td>all applications (self-estimated data qualities) from ( M )</td>
</tr>
<tr>
<td>( D )</td>
<td>all sensing duty allocations to ( M )</td>
<td>( E )</td>
<td>all agents recruited by ( s ) for ( \tau )</td>
</tr>
<tr>
<td>( \hat{O} )</td>
<td>all sensing observations from ( E )</td>
<td>( \hat{Q} )</td>
<td>all actual data quality indications of ( \hat{E} )</td>
</tr>
<tr>
<td>( o_\tau )</td>
<td>the discovered sensing truth (result of ( \tau ))</td>
<td>( \hat{q}_i )</td>
<td>the self-estimated data quality of ( \tau )'s employee ( i )</td>
</tr>
<tr>
<td>( o_i )</td>
<td>the original sensing observation of ( i )</td>
<td>( \hat{o}_i )</td>
<td>the reported observation of ( i )</td>
</tr>
<tr>
<td>( b_i )</td>
<td>the intrinsic sensing bias of ( i )</td>
<td>( d_i )</td>
<td>the allocated sensing duty of ( i )</td>
</tr>
<tr>
<td>( q_i )</td>
<td>the actual data quality indication of ( i )</td>
<td>( u_i )</td>
<td>the gross sensing profit of ( i )</td>
</tr>
<tr>
<td>( p_i )</td>
<td>the net sensing profit of ( i )</td>
<td>( \delta_i )</td>
<td>the relative contribution of ( i )</td>
</tr>
</tbody>
</table>

1. The physical information requested by a specific MCS application, e.g. outdoor temperature, noise level, and air quality within a specific geographical area.
2. All agent identities referred in the rest of the paper are pseudo-identities.
3. The unit of \( B \) is application-specified, which can be monetary unit (e.g. cash back in $), transferable credit unit (e.g. free cellular data traffic in MB), etc.
4. Considering the diversity of application-specified data volume measurements, for the ease of subsequent sensing profit modeling, we measure the duty of each employee from the perspective of sensing budget: assuming that the CSP maintains a fixed mapping from data/unit (e.g. MB) to budget/unit (e.g. $), then duty \( d_i \) denotes that agent \( i \) should report data worth \( d_i \) units of budget.

Fig. 1: The general architecture of privacy-preserving MCSs.
3) Finally, with the established methods for both effective incentivization and truthful reporting, we assume that each agent $i \in \mathcal{M}$ does not know the actual data quality indication $q_i$ (e.g. because of the unknown sensing bias of each agent in practice), but is willing to apply to participate in $\tau$ with a more accurate $\hat{q}_i$ (ideally $\hat{q}_i = q_i$) for a higher profit. In this case, the problem needs to be solved to guarantee the accuracy of their own data quality by themselves.

4 SOLUTION OVERVIEW

In this section, we provide an overview of our solution based on the ‘divide and conquer’ methodology to the aforementioned sub-problems, which is illustrated in Fig.2.

Considering the most intense assumption of sub-problem 1) in Subsection 3.2 (i.e. each $i \in \mathcal{M}$ knows $q_i$ and applies with $\hat{q}_i = q_i$), we develop an optimal sensing duty allocation method for $s$ in Section 5. With the proposed method, for a specific task $\tau$, $s$ can selectively recruit applicants with high-quality data to guarantee the accuracy of MCS services, and can guarantee that each employed agent receives the maximum sensing profit for effective incentivization without exceeding the total sensing budget limitation.

Then, to consider more general scenarios in practice, the assumption of truthful application is loosened in sub-problem 2) in Subsection 3.2 (i.e. each $i \in \mathcal{M}$ knows $q_i$ but may apply with $\hat{q}_i \neq q_i$). Since the method in Section 5 alone cannot handle the impact of untruthful applications, we develop a supplementary method based on algorithmic mechanism design for $s$ in Section 6. With the proposed method, to get the maximum sensing profit, all agents need to apply truthfully (i.e. $\hat{q}_i = q_i$) because of the introduction of the agent social cost from the data quality perspective (in Subsection 6.3).

Finally, for the most general scenario where agents do not know their own data quality (i.e. the assumption of sub-problem 3) in Subsection 3.2), applicants will not have high sensing profits unless they can accurately estimate their data quality by themselves (i.e. agent $i$ has $\hat{q}_i$ as close to $q_i$ as possible). Considering this, we develop an adaptive data quality self-estimation method in Section 7. With the proposed method, all agents can actively calibrate their own potential sensing bias and enhance the accuracy of their own data quality self-estimation based on the feedback from $s$.

In the following sections, we introduce each of these methods in detail.

5 QUALITY-DRIVEN AGENT RECRUITMENT AND SENSING DUTY ALLOCATION

In this section, we assume that each agent $i \in \mathcal{M}$ knows $i$’s actual data quality indication $q_i$ and truthfully applies to participate in MCS tasks with $\hat{q}_i = q_i$.

In this case, to guarantee the accuracy of the MCS service, we develop a method for $s$ to recruit agents with higher data quality under limited budget. Meanwhile, an optimal sensing duty allocation is conducted to maximize the profit of each employee as an effective incentivization.

Problem Segmentation

1) assuming that agent applies with $\hat{q}_i = q_i$.
2) assuming that agent possesses accurate $q_i$.
3) assuming that agent wants $\hat{q}_i$ closer to $q_i$.

Solution Segmentation

5.1 Agent Gross Sensing Profit

For a single task $\tau$, we define the quality-driven gross sensing profit of each agent $i \in \mathcal{E}$ as:

$$u_i = \frac{q_id_i}{\sum_{j \in \mathcal{E}} q_jd_j} B - \frac{d_i}{q_i}$$

where $\mathcal{E}$ is the set of all employees (and $|\mathcal{E}|$), $B$ is the limited total budget for $\tau$ determined by the service provider $s$, $q_i$ is the data quality indication of agent $i$, and $d_i$ is $i$’s allocated duty (the determination of $\mathcal{E}$, $q_i$ and $d_i$ will be explicitly discussed later). Actually, the first item of the Right-Hand-Side (RHS) of Eq.(1) reflects the utility of $i$’s reported data comparing to that of all other employees, and the second item reflects $i$’s cost for accomplishing the allocated sensing duty $5$.

According to Eq.(1), the first-order derivative of $u_i$ on $q_i$ is:

$$\frac{\partial u_i}{\partial q_i} = \frac{d_iB \sum_{j \in \mathcal{E}\{i\}} q_jd_j}{(\sum_{j \in \mathcal{E}} q_jd_j)^2} + \frac{d_i}{q_i^2} > 0,$$

where $\mathcal{E}\{i\}$ denotes all employees except for $i$. Therefore, $u_i$ is a monotonically increasing function of $q_i$ (with a fixed set $\mathcal{D} = \{d_1, d_2, ..., d_{|\mathcal{E}|}\}$).

Similarly, we have the first-order derivative of $u_i$ on $d_i$ as:

$$\frac{\partial u_i}{\partial d_i} = \frac{q_iB}{\sum_{j \in \mathcal{E}} q_jd_j} - \frac{q_i^2d_iB}{(\sum_{j \in \mathcal{E}} q_jd_j)^2} - \frac{1}{q_i},$$

and the second-order derivative as:

$$\frac{\partial^2 u_i}{\partial d_i^2} = -\frac{2q_i^2B \sum_{j \in \mathcal{E}\{i\}} q_jd_j}{(\sum_{j \in \mathcal{E}} q_jd_j)^3} < 0.$$

Therefore, with a fixed set $\mathcal{Q} = \{q_1, q_2, ..., q_{|\mathcal{E}|}\}$, for each $i \in \mathcal{E}$, there exists a unique maximum $u_i^*$, and the corresponding

5. We model the cost as $d_i/q_i$ to reflect the impact of data quality on the cost-effectiveness for agents to participate in MCS tasks. Since our focus is the quality of MCS data, we intuitively assume that it consumes the same amount of resource $B$ (CPU cycle, energy, bandwidth, etc.) for each agent $i$ to accomplish the same amount of duty $d$, which, however, provides $s$ with a different amount of utility $q_i$. In this case, $R(q,d)$ reflects $i$’s cost for providing with one unit of utility, which can be simplified as $1/q_i$ by setting $R = d$ (i.e. the simplest linear relation between consumed resource and accomplished duty for the ease of modeling. Other specific resource limitations in practice can be integrated without affecting the effectiveness of our model). Then, from $i$’s perspective, providing $s$ with utility $q_i.d$ is equal to accomplishing duty $d$ since $q_i$ is fixed for $i$. Therefore, we model $i$’s cost for accomplishing the allocated duty $d_i$ as $d_i/q_i$. 

duty allocation \( d_i^* \) (if exists) could be calculated by solving \( \partial u_i / \partial d_i = 0 \), as:

\[
d_i^* = \sqrt{B \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j - \frac{1}{q_i} \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j},
\]

(5)

5.2 The Optimal Sensing Duty Allocation

Intuitively, for the optimal duty allocation, we want to set \( d_i = d_i^* \) for \( \forall i \in \mathcal{E} \) of task \( \tau \) for effective incentivization. However, according to Eq.(5), the optimal allocation for each employee is constrained by the total budget, the data quality, and the duty allocation of all other employees.

For \( s_i \), it is desirable to determine the optimal duty allocation only based on the participating applications (i.e. the data quality indications). To achieve this, we conduct the following transformation inspired by [32].

According to Eq.(5), there is:

\[
q_i d_i = q_i \sqrt{B \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j - \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j},
\]

(6)

then, by moving the second item of the RHS of Eq.(6) to the left and solving for \( \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j \), we have:

\[
\sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j = \left( \frac{\sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j}{q_i^2 B} \right)^2.
\]

(7)

Insert Eq.(7) into Eq.(6) and sum up \( q_i d_i \) for \( \forall i \in \mathcal{E} \), there is:

\[
\sum_{i \in \mathcal{E}} q_j d_j = E \sum_{i \in \mathcal{E}} q_j d_j - \sum_{i \in \mathcal{E}} \left( \frac{\sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j}{q_i^2 B} \right)^2,
\]

where \( E = |\mathcal{E}| \). Then, by solving Eq.(8) for \( \sum_{j \in \mathcal{E}} q_j d_j \), we have:

\[
\sum_{j \in \mathcal{E}} q_j d_j = \frac{(E - 1)B}{\sum_{j \in \mathcal{E}} q_j^2}.
\]

(9)

Insert Eq.(9) into Eq.(7), there is:

\[
\sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j = \frac{(E - 1)^2 B}{\sum_{j \in \mathcal{E} \setminus \{i\}} q_j^2}.
\]

(10)

Since \( q_i d_i = \sum_{j \in \mathcal{E}} q_j d_j - \sum_{j \in \mathcal{E} \setminus \{i\}} q_j d_j \), by inserting Eq.(9) and (10) and solving for \( d_i \), there is:

\[
d_i = \frac{(E - 1)B}{q_i \sum_{j \in \mathcal{E}} q_j^2} \left( 1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2} \right).
\]

(11)

Insert Eq.(11) into Eq.(1), there is:

\[
u_i = \left( 1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2} \right)^2 B.
\]

(12)

Considering Eq.(11), it is viable for \( s_i \) to determine the optimal duty allocation for all applicants only based on their data quality indications. Based on this, we propose the Quality-Driven Sensing Duty Allocation algorithm (i.e. Algorithm 1), which determines the task employee selection and calculates the optimal duty allocation simultaneously. In general, Algorithm 1 intends to recruit as many applicants as possible in the descending order of their data quality with the limited total budget.

For the practicability discussion of Algorithm 1, we have the following proposition.

**Algorithm 1: Quality-Driven Sensing Duty Allocation**

**Input:**

- \( Q = \{q_1, q_2, ..., q_M\} \): data quality indications of \( \forall i \in \mathcal{M} \);
- \( B \): the total budget for the current task \( \tau \);

**Output:**

- \( \mathcal{E} \): the set of employees recruited by task \( \tau \);
- \( \mathcal{D} = \{d_1, d_2, ..., d_M\} \): duty allocations of \( \forall i \in \mathcal{M} \).

1. put \( \forall q_i \in Q \) in descending order;
2. store the agent IDs in the sorted order as vector \( s \);
3. set \( s_i \) as the \( i \)'s tuple of \( s \);
4. \( D \leftarrow \emptyset \), \( \mathcal{E} \leftarrow \{s_1\}, i \leftarrow 2 \);
5. while \( i \leq M \) & \& \( q_{s_i} > \sqrt{\frac{E}{q_{s_i}^2 + \sum_{j \in \mathcal{E}} q_j^2}} \) do
6. \( \mathcal{E} \leftarrow \mathcal{E} \cup \{s_i\} \);
7. \( i++ \);
8. for \( \forall i \in \mathcal{M} \) do
9. if \( i \in \mathcal{E} \) then
10. \( d_i = \frac{(E-1)B}{q_i \sum_{j \in \mathcal{E}} q_j^2} \left( 1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2} \right) \);
11. else
12. \( d_i = 0 \);
13. \( D \leftarrow D \cup \{d_i\} \);
14 return \( \mathcal{E}, D \). 

**Proposition 1.** The duty allocation \( \mathcal{D} = \{d_1, d_2, ..., d_M\} \) from Algorithm 1 fulfills following requirements:

1. Positive duty: for \( \forall i \in \mathcal{E}, d_i > 0 \);
2. Personal rationality: for \( \forall i \in \mathcal{E}, u_i > 0 \);
3. Budget limitation: for any specific \( \mathcal{D}, \sum_{i \in \mathcal{M}} u_i < B \);
4. Optimal allocation: with given \( Q \) and \( B \), each \( d_i \in \mathcal{D} \) is the optimal allocation for the corresponding \( i \in \mathcal{M} \).

**Proof.** We prove each of the sub-propositions sequentially:

1. The positive duty requirement indicates that all agents employed by Algorithm 1 should be allocated with positive sensing duties in practice. According to Algorithm 1, there are:

\[
u_{s_1} \geq \nu_{s_2} \geq \cdots \geq \nu_{s_M}.
\]

(from lines 1 to 3 of Algorithm 1)

and:

\[
u_{s_M} > \sqrt{\frac{E - 1}{\sum_{j \in \mathcal{E}} q_j^2}}.
\]

(from line 5 of Algorithm 1)

Therefore, for \( \forall i \in \mathcal{E} \), there is:

\[
u_i > \sqrt{\frac{E - 1}{\sum_{j \in \mathcal{E}} q_j^2}}.
\]

(from Eq.(13) and (14))

By inserting Eq.(15) into Eq.(11), there is \( d_i > 0 \) for \( \forall i \in \mathcal{E} \). In this case, we prove that all agents employed by Algorithm 1 are allocated with positive sensing duties.

2. The personal rationality requirements indicates that all agents employed by Algorithm 1 should have
positive gross sensing profits in practice. Since for \( \forall i \in \mathcal{E}, d_i > 0 \), according to Eq.(11), there is:

\[
1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2} > 0. \tag{16}
\]

Insert Eq.(16) into Eq.(12), there is \( u_i > 0 \) for \( \forall i \in \mathcal{E} \). In this case, we prove that all agents employed by Algorithm 1 are guaranteed with positive gross sensing profits.

3) The budget limitation requirement indicates that the total gross sensing profit of all agents should not exceed the total sensing budget. Considering the total gross sensing profit of all agents \( i \) in \( \mathcal{M} \), there is:

\[
\sum_{i \in \mathcal{E}} u_i = \sum_{i \in \mathcal{E}} u_i + \sum_{i \in \mathcal{M} \setminus \mathcal{E}} u_i \\
= B \sum_{i \in \mathcal{E}} (1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2})^2 + 0
\]

(from Eq.(12) and Algorithm 1)

\[
< B(\sum_{i \in \mathcal{E}} (1 - \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2}))^2
\]

(from 1 - \( \frac{E - 1}{q_i^2 \sum_{j \in \mathcal{E}} q_j^2} > 0 \) for \( \forall i \in \mathcal{E} \))

\[
= B(E - \sum_{i \in \mathcal{E}} q_i^2 \sum_{j \in \mathcal{E}} q_j^2)^2
\]

\[
= B(E - \sum_{j \in \mathcal{E}} q_j^2 \sum_{i \in \mathcal{E}} q_i^2)^2
\]

\[
= B.
\tag{17}
\]

In this case, we prove that any duty allocation from Algorithm 1 guarantees that the total gross sensing profit of all agents does not exceed the total sensing budget.

4) The optimal allocation requirement indicates that the duty allocation from Algorithm 1 should be the best allocation to all applicants. We discuss the optimality of the duty allocation for applicants who are unemployed and employed by task \( \tau \), respectively. On one hand, for any \( i \in \mathcal{M} \setminus \mathcal{E} \) (i.e. unemployed), by setting \( \mathcal{E}' = \mathcal{E} \cup \{i\} \), there is:

\[
q_i \leq \frac{E}{\sqrt{q_i^2 + \sum_{j \in \mathcal{E}} q_j^2}}
\]

(from line 5 of Algorithm 1) \tag{18}

\[
= \sqrt{\frac{1}{\sum_{j \in \mathcal{E} \setminus \{i\}} q_j^2} - 1}
\]

By inserting Eq.(18) into Eq.(11), there is:

\[
d_i^* = \frac{(|\mathcal{E}'| - 1)B}{q_i^2 \sum_{j \in \mathcal{E} \setminus \{i\}} q_j^2} (1 - \frac{|\mathcal{E}'| - 1}{q_i^2 \sum_{j \in \mathcal{E} \setminus \{i\}} q_j^2}) \leq 0. \tag{19}
\]

Therefore, considering Eq.(1), Eq.(4) and the positive duty constrain of the CSP (i.e. all employed agents have to be allocated with positive duties), the optimal allocation for \( i \) is \( d_i = 0 \), i.e. \( i \) does not participate in \( \tau \), since otherwise \( i \) will have a negative sensing profit. On the other hand, for any \( i \in \mathcal{E} \) (i.e. employed) with the allocated duty \( d_i \), from Algorithm 1, by inserting Eq.(9) and Eq.(11) into Eq.(3), there is:

\[
\frac{\partial u_i}{\partial d_i} = 0. \tag{20}
\]

Therefore, there is \( d_i = d_i^* \) for \( \forall i \in \mathcal{E} \). In summary, we prove that each \( d_i \) in \( \mathcal{D} \) from Algorithm 1 is the optimal duty allocation for the corresponding applicant \( i \) in \( \mathcal{M} \).

Until now, we prove that Algorithm 1 fulfills all requirements in Proposition 1. \( \square \)

According to the analysis above, with Algorithm 1, \( s \) can determine the optimal sensing duty allocation for all applicants under a limited budget. However, the profit of employees could be maximized only if all of the data quality indications were truthfully reported, which is difficult to guarantee considering the self-interested agents in the privacy-preserving scenario (to be discussed later).

### 6 Mechanism Design against Untruthful Reporting

In this section, we loose the assumption in Section 5 as that each agent \( i \in \mathcal{M} \) knows \( i \)'s actual data quality indication \( q_i \), but may apply to participate in MCS tasks with \( \hat{q}_i \neq q_i \) for a potentially higher profit.

In this case, we develop a method based on algorithmic mechanism design [33] to enforce \( i \) applying truthfully with the actual data quality indication \( q_i \) (i.e. \( \hat{q}_i = q_i \)) for the maximum profit.

For a task \( \tau \) with budget \( B \), we denote the actual data quality indications of all \( i \in \mathcal{M} \) as \( q = (q_1, q_2, \ldots, q_M) \), where each \( q_i \) is privately known to \( i \). According to Eq.(2), for agent \( i \), with given \( B, D \) and \( Q \setminus \{q_i\} \), the higher \( q_i \), the higher \( u_i \). Therefore, considering our aforementioned privacy-preserving setting, it is possible that any \( i \in \mathcal{M} \) will claim a fake \( \hat{q}_i \) for a higher \( u_i \).

Define the strategy of all \( i \in \mathcal{M} \) as the self-estimated data quality indications \( \hat{q} = (\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_M) \) they claim, where \( \hat{q}_i \) is not necessarily the same as \( q_i \), and it can be deliberately chosen by \( i \) to maximize the profit. Actually, the determination of the claiming strategies can be modeled as a non-cooperative game with incomplete information among all \( i \in \mathcal{M} \) and \( s \). To enforce truthful reporting, we develop a Vickery-Clarks-Groves (VCG) mechanism [34] as follows.

#### 6.1 Mechanism Setups

Consider all \( i \in \mathcal{M} \) as \( M \) players; define \( \mathcal{D} \) as the set of all possible outcomes (i.e. sensing duty allocations); define \( \mathcal{Q} \) as the set of all possible strategies (i.e. claimed quality indications); set the valuation function of each \( i \in \mathcal{M} \) as \( u_i(q, d) \equiv u_i(q, d) \) (i.e. Eq.(1)), where \( q \in \mathcal{Q} \) and \( d \in \mathcal{D} \).
6.2 Allocation Rule

Define the value of the outcome function \( f(u_q) \) as the social efficient allocation under strategy \( q^* \):

\[
 f(u_q) \triangleq \arg \max_{d \in D} \sum_{i \in M} u_{i,q}(d),
\]

where we call \( f(u_q) \) is efficient since:

\[
 \sum_{i \in M} u_{i,q}(f(u_q)) \geq \sum_{i \in M} u_{i,q}(d)
\]

for \( \forall q \in Q \) and \( \forall d \in D \).

Let \( d^* = (d_1, d_2, \ldots, d_M) \in D \) be the duty allocation determined by Algorithm 1 (with given \( \hat{q} \) and \( B \)), we have the following proposition.

**Proposition 2.** For the duty allocation \( d^*(\hat{q}) \) determined by Algorithm 1, there is:

\[
 d^*(\hat{q}) = f(u_q).
\]

**Proof.** According to Algorithm 1, there is:

\[
 \sum_{i \in M} u_{i,q}(d^*) = \sum_{i \in M} u_{i,q}(d^*)
\]

and, for each \( i \in E \):

\[
 d^*_i(\hat{q}) = \arg \max_{d \in D} u_{i,q}(d),
\]

where \( x_{(i)} \) denotes the \( i \)th entry of vector \( x \). Considering Equ.(3) and Eq.(4), for each \( i \in E \), \( u_{i,q} \) is a strictly concave function of \( d_i > 0 \). Meanwhile, with given \( \hat{q} \), all entries of \( d^*(\hat{q}) \) are mutually independent according to Eq.(11).

Therefore, we have:

\[
 d^*(\hat{q}) = \arg \max_{d \in D} \sum_{i \in M} u_{i,q}(d) = f(u_q),
\]

i.e. the optimal allocation from Algorithm 1 is an efficient social allocation for \( \forall i \in M \) under strategy \( \hat{q} \).

Considering the discussion above, we set \( f(u_q) = d^*(\hat{q}) \) from Algorithm 1.

6.3 Payment Rule

To enforce truthful reporting, we define agent \( i \)‘s net sensing profit \( p_i \) for participating task \( \tau \) as:

\[
 p_i = u_{i,q}(f(u_q)) - c_i,
\]

where \( u_{i,q}(f(u_q)) \) denotes the gross sensing utility calculated using Eq.(1) with all actual data quality indications \( q \) and the duty allocation \( f(u_q) \) from Algorithm 1 based on strategy \( q \), and \( c_i \) denotes \( i \)’s social cost under strategy \( \hat{q} \):

\[
 c_i = \sum_{j \in M \setminus \{i\}} u_{j,q_{-i}}(f(u_{q_{-i}})) - \sum_{j \in M \setminus \{i\}} u_{j,q}(f(u_q)),
\]

where \( q_{-i} \) and \( q_{-i} \) represent the actual and claimed (i.e. self-estimated) data quality indications of all agents in \( M \) except for \( i \), respectively. The first item of Eq.(28)’s RHS represents the maximum social gross profit that all \( j \in M \setminus \{i\} \) could have (without \( i \)’s existence), while the following item represents the social gross profit that all \( j \in M \setminus \{i\} \) have when \( f(u_q) \) is adopted. Actually, for \( \forall i \in M \), \( c_i \) indicates \( i \)‘s impact on the current task \( \tau \) (i.e. \( c_i > 0 \): hindering, \( c_i < 0 \): contributing, \( c_i = 0 \): not affecting).

With the definitions above, we have the following proposition.

**Proposition 3.** Under the allocation rule (Eq.(21)) and the payment rule (Eq.(27), Eq.(28)), the best strategy for \( \forall i \in M \) to truthfully report their data quality indications \( (\hat{q}^* = q) \).

**Proof.** Assuming that any agent \( i \in M \) deliberately reports a fake \( q_i' \neq q_i \) for \( s \) while all others are reporting truthfully. According to Eq.(1), (21), (27) and (28), \( i \)’s net profit by untruthful reporting is:

\[
 p'_i = \sum_{j \in M} u_{j,q}(f(u_{q_{j-1}})) - \max_{d \in D} \sum_{j \in M \setminus \{i\}} u_{j,q_{-i}}(d),
\]

and \( i \)’s net profit by truthful reporting is:

\[
 p_i = \max_{d \in D} \sum_{j \in M} u_{j,q}(d) - \max_{d \in D} \sum_{j \in M \setminus \{i\}} u_{j,q_{-i}}(d).
\]

Since:

\[
 \sum_{j \in M} u_{j,q}(f(u_{q_{j-1}})) \leq \max_{d \in D} \sum_{j \in M} u_{j,q}(d),
\]

we have \( p'_i \leq p_i \) for \( \forall i \in M \), and the best strategy for all applying agents is to report truthfully (i.e. \( \hat{q}^* = q \)).

According to the analysis above, with the designed allocation and payment rules, all agents in \( M \) can achieve the maximum profit only when they apply for the participation using accurate self-estimated data quality indications and report truthfully during the sensing task. However, in practice, agents do not have their actual data quality indications because of, for example, the potential sensing bias of mobile device that is device-specified and unknown to the device owner, and it is challenging for them to achieve accurate data quality self-estimations in the privacy-preserving scenarios (to be discussed later).

7 Adaptive Data Quality Self-estimation

In this section, we consider more practical scenarios where each agent \( i \in N \) does not know \( i \)’s actual data quality indication \( q_i \) but is willing to participate in MCS tasks.

With the adoption of mechanisms in Section 5 and 6, each agent \( i \in N \) needs to continuously enhance the accuracy of \( \hat{q}_i \) (ideally \( \hat{q}_i = q_i \)) to get a higher net sensing profit \( p_i \). To achieve this, we concentrate on two fundamental problems in this section:

1) For \( s \), how to evaluate the quality of sensing observations reported by the agents?
2) For each agent \( i \in N \), how to locally estimate and enhance the quality of sensing observations?

7.1 Quality Evaluation of Reported Observations

For task \( \tau \), let agent \( i \in E \) report an observation \( o_{i,0} \) to \( s \). Assuming that the ground truth of \( \tau \) is \( o_\tau \) (to be discussed later), we define \( i \)’s actual data quality indication \( q_i \) as:

\[
 q_i = e^{-\alpha(o_{i,0} - o_\tau)},
\]

where \( \alpha \) is the learning rate.

6. For simplicity, we assume that the reported observation of each agent \( i \) is the average of all \( i \)’s sensing readings during the current task.
7.2 Self-estimation of the Data Quality

According to Eq.(27) and (28), for each $i \in \mathcal{E}$, the more accurate $\hat{q}_i$ is (i.e. the smaller the difference between $\hat{q}_i$ and $q_i$ is), the higher $p_i$ is. All agents need a method to actively enhance their data quality by conducting accurate data quality self-estimation.

7.2.1 Agent Sensing Bias Calibration

Intuitively, we set that the original observation $o_i$ of each agent $i \in \mathcal{N}$ has an unknown but relatively stable sensing bias $b_i$, which should be inherently determined by the operating habit of the mobile user or the hardware limitation of the mobile device.

For $\forall i \in \mathcal{N}$, $b_i$ is maintained by agent $i$ and iteratively calibrated after each task. Ideally, $i$ should be able to achieve $\hat{o}_i = o_i - b_i = o_\tau$. Specifically, for each $i \in \mathcal{E}$ of $\tau$, after receiving the relative contribution $\delta_i$ and $\tau$’s sensing truth from $s$, the update rule of $b_i$ is defined as:

$$b_i = \begin{cases} 
\| \delta_i \| b_i' + (1 - \| \delta_i \|) (o_i - o_\tau) & \text{if } \delta_i \geq 0, \\
(1 - \| \delta_i \|) b_i' + |\delta_i| (o_i - o_\tau) & \text{if } \delta_i < 0, 
\end{cases}$$  \hspace{1cm} (33)

where $b_i'$ is $i$’s previous bias before participating in task $\tau$.

According to Eq.(33), when $\delta_i \geq 0$, agent $i$ is contributing no less than the average, and the higher $\delta_i$ is (i.e. $\hat{o}_i$ is closer to $o_\tau$), the more $i$ will trust the self-maintained bias $b_i$ during the update process; when $\delta_i < 0$, $b_i$ will be moved towards the actual difference between the original observation $o_i$ and the discovered truth $o_\tau$ (i.e. $\hat{o}_i - o_\tau$).

7.2.2 Agent Data Quality Self-Estimation

Similar to the agent bias, each agent $i \in \mathcal{N}$ maintains a self-estimated data quality indication $\hat{q}_i$, which is updated after each task. Specifically, for each employee $i$ of task $\tau$, $\hat{q}_i$ is calculated as:

$$\hat{q}_i = e^{-\alpha(\hat{o}_i - o_\tau)};$$  \hspace{1cm} (34)

where $\alpha$ is $i$’s reported observation and the sensing truth of $i$’s last task, and $0 < \gamma < 1$ determines the impact of the current task $\tau$ on the update of $\hat{q}_i$ (i.e. concentrate more on the previous experience or the current result).

According to the analysis above, with the designed data quality self-estimation method, all agents in $\mathcal{N}$ can actively pursue higher net sensing profits by gradually enhancing the accuracy of their self-estimated data quality indications.

Until now, by combining our methods in Sections 5 to 7, we have developed a systematic solution for guaranteeing the quality of reported data in privacy-preserving MCSs.

8 PERFORMANCE EVALUATION

In this section, we conducted extensive simulations based on real-world MCS traces to evaluate the correctness and effectiveness of our design.

8.1 Experimental Methodology

To evaluate the performance of our mechanism in practical systems, we constructed a large-scale MCS simulator based on the OMNeT++ 4.6 platform [36] with the implementation of the mechanism proposed in Section 5, 6, and 7. Based on the MCS, we simulated the process of the crowd-sourced outdoor temperature sensing task 7 using real-world

7. It should be better for us to select MCS use cases focusing on more privacy-sensitive data (personal medical data, personal financial data, etc.) to demonstrate the effectiveness of our solution in real-world applications. However, because of the difficulty of accessing such privacy-sensitive data in practice, we use real-world MCS data (i.e. crowd-sourced temperature) that are available in public with manually introduced untruthful behaviors for our simulations. One thing should be noted is that such a choice does not affect the performance of our solution, which can be directly applied to other practical use cases of privacy-sensitive sensing.
MCS data traces, and evaluated our mechanism’s capability of conducting accurate sensing against untruthful reporting.

8.1.1 Simulation Data-set
In our simulations, we adopted the real-world outdoor temperature sensing traces provided by [37]. Specifically, the data-set used contains 5030 sensing entries opportunistically reported by 367 taxis across the Rome city within 24 hours. Each entry contains a time stamp, an agent identity, a set of location coordinates (latitude, longitude), and a temperature observation.

8.1.2 Simulator Structure
According to the data-set, we constructed an MCS composed of one service provider \(s\) and 367 participating agents with a \(c^{++}\) implementation of our proposed mechanism. Once a simulation round is started, \(s\) continuously publishes a new sensing task with a random (integer) interval from 0 to 2 simulation seconds. For a specific task \(\tau\) started at time point \(t\), \(s\) generates a random desired sensing location \(L_\tau\) within the maximum sensing area (determined by the data-set). According to the data-set, any agent, within the effective sensing area (i.e., the circular area with \(L_\tau\) as the center, and an effective range \(r_\tau\) determined by \(s\) as the radius, see Fig.1), who possesses an observation collected within the time range of \(t \pm 60\) simulation seconds automatically applies to participate in \(\tau\). One thing should be noted is that: to make the most of the data-set, we expanded each round of the simulation from 86400 to 259200 simulation seconds, and each time point \(t\) during the simulation was mapped to a random time point in the actual data-set.

8.1.3 Simulated Threats
To thoroughly mimic the potential threats in practical MCSs, we defined 3 different kinds of agent behaviors in each task as follows:

- **Random Observation (RO):** the selected agent reports a random observation \(^8\) to \(s\);
- **Untruthful Application (UA):** the selected agent applies for participation using a higher (but fake) self-estimated data quality indication;
- **Regulated Behavior (RB):** the selected agent behaves benignly according to all of the designed regulations.

8. A random temperature within \([2^\circ C, 24^\circ C]\) according to the data-set.

Based on these definitions, we conducted extensive simulations considering different practical scenarios, which will be explicitly discussed later.

8.1.4 Parameter Settings
During the simulations, we set: the effective sensing range \(r_\tau = 2KM\), the least participant number of a task \(E_i = 4\), the initial bias of all agents \(b_{init} = 0\), the initial self-estimated data quality indication of all agents \(q_{init} = 1\), the total budget of each task \(B = 1\), the truth discovery threshold \(\epsilon = 0.1\), the quality dropping sensitivity \(\alpha = 1\), the contribution dropping sensitivity \(\beta = 6\), and the quality self-estimation update sensitivity \(\gamma = 0.1\). Particularly, we conducted an empirical study on the determination of \(\alpha, \beta, \) and \(\gamma\) in Subsection 8.2.

8.2 Simulation Scenarios, Results and Discussions
In these simulations, we first evaluated the effectiveness of the proposed mechanism on providing accurate sensing services. Then, we studied the performance of our mechanism against general untruthful reporting with different intensities. After that, we investigated the impact of untruthful reporting on the sensing profit of agents. Finally, we discussed the issue of parameter determinations.

8.2.1 The Effectiveness of the Proposed Mechanism
For effectiveness evaluation, we conducted one round of simulation within each of the following scenarios: the original scenario (i.e. ‘org’) where all agents’ original observations were directly reported to the truth discovery process, the processed scenario (i.e. ‘prc’) where the entire proposed mechanism was adopted, and the selected scenario (i.e. ‘slc’) where representative agents were selected to directly report their original observations while all others reporting with the entire proposed mechanism. All agents behaved as regulated within all scenarios.

Fig.3 shows the cumulative distributions of the discovered truth (i.e. ‘dt’) and the standard deviation (i.e. ‘rstd’) of all observations reported to \(s\) in each task within the original and the processed scenarios. According to Fig.3a, we can see that prc-dts (11.69°C in average) and org-dts (11.77°C in average) basically share the same distribution, but the distribution of prc-dts converges faster at specific temperatures (as marked) than that of org-dts. Meanwhile, the rstds within two scenarios are quite different according to Fig.3b, where

---

Fig. 3: Effectiveness on sensing accuracy.

Fig. 4: Effectiveness on sensing bias calibration.
TABLE 2: Agent net profits with/without bias calibration.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>with BC</th>
<th>without BC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Task Count</td>
</tr>
<tr>
<td>Agent 358</td>
<td>0.21</td>
<td>805</td>
</tr>
<tr>
<td>Agent 266</td>
<td>0.21</td>
<td>210</td>
</tr>
</tbody>
</table>

The probability distributions of the original net profits with/without bias calibration are as follows:

- Agent 358: Net profit probability distribution
- Agent 366: Net profit probability distribution

Fig. 5: Effectiveness on quality-based profit enhancement.

$prc-stds$ (0.44°C in average) are obviously lower than $org-stds$ (1.00°C in average).

To evaluate the effectiveness of our mechanism on enhancing the data quality of participating agents, we selected two representative agents (i.e. the $TopCount$ agent 358, and the $TopBias$ agent 266) as examples. Fig.4 shows the variation of the sensing errors (i.e. the actual difference between the discovered truth and the reported observation) of each selected agent along time within the original and the processed scenarios (i.e. ‘org-err’ and ‘prc-err’), as well as the self-maintained sensing bias of each agent within the processed scenario (i.e. ‘self-bias’). As shown, for both agents, $prc-errs$ gradually decline to become relatively stable values around 0 (i.e. around 0.02°C for agent 358, and around −0.05°C for agent 266) after the fluctuations at the beginning of the simulation, which are quite different from $org-errs$ (i.e. −0.34°C in average with a 0.22°C std for agent 358, and 2.36°C in average with a 0.93°C std for agent 266). Besides, it is obvious that both agents manage to determine their relatively stable inherent sensing biases (i.e. around 0.06°C for agent 358, and around 2.48°C for agent 266) after multiple participations, as we expected in Subsection 7.2.

Still regarding agents 358 and 266 as representatives, Fig.5 depicts the cumulative distributions of their net profits (i.e. ‘np’) from each task within the processed and the selected scenarios (i.e. ‘prc-np’ and ‘sle-np’). As shown, for both agents, compared to the $nps$ in the selected scenario (where agent 358 and 266 were selected not to calibrate their sensing biases), their $nps$ increase significantly when our mechanism is adopted (see Table 2).

As demonstrated, our mechanism manages to effectively reduce the controversies among reported observations without disrupting the basic distribution of the crowdsensing results. Meanwhile, it enables the participating agents to actively enhance the quality of their sensing observations (and to get higher sensing profits).

8.2.2 The Impact of General Untruthful Reporting on the Sensing Accuracy

To evaluate the performance of the proposed mechanism where untruthful reporting is prevalent, we specifically consider 3 types of practical untruthful behaviors from the agents: RO, RO+UA, and UA (see Subsection 8.1), and conducted 3 sets of simulations where each type of the untruthful behaviors was introduced, respectively. For each set of simulations, different numbers (i.e. 10, 20, 30, and 40) of the participating agents were selected randomly as ‘malicious’ who reported untruthfully. For each round of simulations, the discovered truth (i.e. ‘dt’) and the standard deviation (i.e. ‘rstd’) of all reported observations of each task were recorded. The results where all agents reported truthfully (i.e. RB) were treated as the baseline (i.e. ‘base’).

Fig.6 shows the cumulative distributions of the recorded results when ROs with different intensities were introduced. According to Fig.6a, although the introduced ROs manage to cause limited sensing concept drifts (see Table 3), the basic distribution pattern of the $dts$ with ROs is still preserved (as marked). Meanwhile, all $rstds$ with ROs have declined according to Fig.6b (also see Table 3). Actually, during the simulations, all RO agents are ‘abandoned’ (i.e. not employed, to be demonstrated later) by $s$ quite soon because of their high sensing errors and rapidly declining self-estimated data quality indications. Since the observations reported by other RB agents are not affected, a referable $dt$ distribution can still be extracted without suffering from the external deviations caused by the introduced ROs.

Fig.7 shows the results when RO+UAs with different intensities were introduced. As shown, compared to ROs, RO+UAs have deeper impacts on the cumulative distributions of both the $dts$ and the $rstds$. According to Fig.7a, the basic distribution pattern of the $dts$ is disturbed more obviously (as marked, also see Table 4). Besides, all $rstds$ with RO+UAs have increased according to Fig.7b (also see...
TABLE 4: Statistics of the sensing results with RO+UAs.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Average $dt^\circ C$</th>
<th>Average $rstd^\circ C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>11.69</td>
<td>0.44</td>
</tr>
<tr>
<td>random 10</td>
<td>11.99</td>
<td>1.64</td>
</tr>
<tr>
<td>random 20</td>
<td>12.15</td>
<td>2.16</td>
</tr>
<tr>
<td>random 30</td>
<td>11.77</td>
<td>2.8</td>
</tr>
<tr>
<td>random 40</td>
<td>11.32</td>
<td>3.81</td>
</tr>
</tbody>
</table>

Table 4). Actually, during simulations, RO+UA agents are employed more often because of their higher (but fake) self-estimated data quality indications, therefore external deviations caused by corresponding ROs disturb the distribution pattern of the $dls$. However, one thing should be noted is that our mechanism is designed to punish RO+UA agents by severely reducing their sensing profits (to be demonstrated later), which strictly prevents rational agents from introducing such kind of untruthful behaviors.

Fig.8 shows the results when UAs with different intensities were introduced. As shown, it is clear that UAs have nearly negligible impact on the cumulative distribution of neither the $dls$ nor the $rstds$ (see Table 5). Actually, during the simulations, even that UA agents are employed more often because of their higher (but fake) self-estimated data quality indications, the quality of their reported observations is still guaranteed by the continuous bias calibration process. Moreover, one thing should be noted is that, despite the fact that UAs have negligible impact on the sensing accuracy, our mechanism is designed to prevent UA agents from getting higher profits (to be demonstrated later).

As shown, with our mechanism, neither ROs nor UAs from up to 40 agents (i.e. 10% of all participating agents) lead to obvious impact on the sensing accuracy.

8.2.3 The Impact of Untruthful Reporting on Agents’ Profits

To evaluate the performance of our mechanism in terms of preventing agents from reporting untruthfully (by reducing sensing profits), we selected two sets of representative agents (i.e. each set contains the top 4 of the TopCount/TopBias agents) as ‘malicious’ to introduce different types of untruthful reporting (i.e. RO, RO+UA, and UA) during the simulations. The ‘net profit’ (i.e. ‘np’) of each selected agent for participating in each task was recorded. The results where all agents reported truthfully (i.e. RB) were treated as the baseline.

Fig.9 shows the cumulative distributions of the $np$s of each selected TopCount agent with RBs (i.e. ‘base’), RO+UAs (i.e. ‘ro+ua’), and UAs (i.e. ‘ua’), respectively. Firstly, the results from the RO scenario are not shown since that each selected agent with ROs was employed by $s$ only once during an entire simulation round, and the corresponding $np$ is no more than $-0.46$. Then, according to Fig.9, it is clear that the $np$s of all agents are severely downgraded within the RO+UA scenario. Moreover, with UAs, the $np$s are also downgraded (see Table 6).

According to Fig.10, the impact of untruthful reporting on the selected TopBias agents remains similar: each RO agent was employed once with a $np$ no more than $-3.03$, the $np$s of all RO+UA agents are dramatically downgraded, and the $np$s within the UA scenario are also downgraded (see Table 7).

As demonstrated, with our mechanism, the best strategy for any rational agent is to behave as regulated and report truthfully, since any type of untruthful reporting will lead to obvious sensing profit reductions.

8.2.4 Discussions on parameter determinations

To evaluate the impact of different parameter settings on the performance of our mechanism, focusing on the quality dropping sensitivity $\alpha$ (see Eq.(32)), the contribution dropping sensitivity $\beta$ (see Line 12 of Algorithm 2), and the quality self-estimation update sensitivity $\gamma$ (see Eq.(34)), we conducted multiple sets of simulations to verify the results of our numerical analysis.

According to Eq.(32), $\alpha$ determines the sensitivity of the Data Quality Indication (DQI) on the sensing error. As shown in Fig.11a, a higher $\alpha$ should induce a quicker drop of the DQI when the sensing error increases, while that a lower $\alpha$ should enable a wider range of distinguishable errors. Considering this, we ran 3 rounds of simulations ($\beta = 6, \gamma = 0.1$) with $\alpha = 1/2/4$, respectively, and all of the recorded DQIs are separately depicted in Fig.11b, 11c, and 11d. As shown, DQIs with a higher $\alpha$ increase quicker along the simulation time, while that the total number of accomplished tasks drops as $\alpha$ increases. Actually, this fits for $\alpha$’s numerical property quite well. With a higher $\alpha$ (i.e. $\alpha = 4$), only agents with lower initial errors are repeatedly
employed during the entire simulation round, and their DQIs increase quicker since considerable sensing controversies are eliminated (i.e. other agents and their observations are ‘abandoned’ because of rapidly declining DQIs at the beginning stage). When α is lower (i.e. α = 1), initial DQIs drop slower, and more agents manage to contribute since they have the time and opportunities for the bias calibration process. In practical systems, α should be determined by s to achieve a trade-off between the sensing truth converging speed and the sensing truth objectivity/accuracy depending on specific crowdsensing requirements.

According to Algorithm 2, β determines the sensitivity on distinguishing different relative contributions of participating agents. As shown in Fig.12a (set the least employee number E = 4), a higher β should reduce the range of distinguishable contributions (i.e. [0, 1] when β = 1, and [0, 0.5] when β = 6), while increasing the distinguishability among different contributions. Considering this, we ran 3 rounds of simulations (α = 1, γ = 0.1) with β = 1/3/6, respectively, and all of the recorded agent relative contributions are separately depicted in Fig.12b, 12c, and 12d. As shown, when β = 1, all of the contributions of participating agents are mapped to the relative contributions with a constrained range of [−0.46, 0.89]. As β increases to 6, the range of corresponding relative contributions expands to [−1, 1], which provides with a proper distinguishability. This fits for β’s numerical property quite well. In practical systems, β should be determined by s properly considering the statistical property (e.g. standard deviations) of the sensing observations to enhance the efficiency of the following agent bias calibration process.

According to Eq.(34), γ determines the update sensitivity of the Self-estimated Data Quality Indications (SDQIs) of all participating agents. As shown in Fig.13a, a higher γ should induce an SDQI update process that is more sensitive to the difference between the previous error and the current error (e.g. with a previous error of 0.5°C, when the current error is 0°C, qnew = 0.64 if γ = 0.1, and qnew = 0.78 if γ = 0.5). This should lead to a quicker agent bias calibration process. Considering this, we ran 3 rounds of simulations (α = 1, β = 6) with γ = 0.1/0.3/0.5, respectively, and all of the recorded Report Standard Deviations (RSTDs) are separately depicted in Fig.13b, 13c, and 13d. As shown, RSTDs with all of the 3 settings gradually decline along the simulation time, and, obviously, the declining speed of RSTDs increases as γ increases. This fits for γ’s numerical property quite well. Actually, with a higher γ, only agents with lower initial biases that remain stable are repeatedly employed during the entire simulation. Other agents (as well as their observations) are ‘abandoned’ because of their rapidly declining SDQIs. In practical systems, γ should be determined by s, similar to α, to achieve a trade-off between the sensing truth converging speed and the sensing truth objectivity/accuracy.

### Table 6: Statistics of the net profits of TopCount agents.

<table>
<thead>
<tr>
<th>Agent</th>
<th>358</th>
<th>47</th>
<th>319</th>
<th>174</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average np with RBs</td>
<td>0.21</td>
<td>0.08</td>
<td>0.33</td>
<td>0.09</td>
</tr>
<tr>
<td>Average np with UAs</td>
<td>0.10</td>
<td>0.06</td>
<td>0.15</td>
<td>0.07</td>
</tr>
</tbody>
</table>

### Table 7: Statistics of the net profits of TopBias agents.

<table>
<thead>
<tr>
<th>Agent</th>
<th>266</th>
<th>92</th>
<th>271</th>
<th>315</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average np with RBs</td>
<td>0.23</td>
<td>−0.13</td>
<td>0.54</td>
<td>0.33</td>
</tr>
<tr>
<td>Average np with UAs</td>
<td>0.19</td>
<td>−0.16</td>
<td>0.08</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Fig. 9: Net profits of untruthful TopCount agents.

Fig. 10: Net profits of untruthful TopBias agents.
9 CONCLUSION

Focusing on MCSs that protect the identities of participating individuals and their precise locations, we propose a mechanism that simultaneously: 1) guarantees the quality of data received by the CSP by eliminating the motivation of the (self-interested) participants to report untruthfully, and 2) enhances the quality of data reported by the participants by encouraging active calibration of intrinsic sensing biases for higher profits. With a thorough discussion concerning the issue of data quality in privacy-preserving MCSs, we first develop a quality-driven agent recruitment and sensing duty allocation method that employs as many agents with high quality data as possible within the limited budget, where the profit of each employee is maximized for effective incentivization. Then, to force the agents to apply and report truthfully, we use algorithmic mechanism design to establish a penalisation policy to ensure that behaving truthfully is the optimal strategy for all agents, even when their personal information is well cloaked. After that, we propose an adaptive data quality self-estimation method for each agent to actively enhance the objectivity and accuracy of their sensing observations to achieve higher profits. Finally, we evaluate the performance of our mechanism based on extensive real-world trace-driven simulations. The results demonstrate that our mechanism manages to guarantee the objectivity and accuracy of crowdsensing results, where inexactitudes regarding reported observations are significantly reduced. More importantly, the adoption of our mechanism not only eliminates the motivation of untruthful behaviors but also stimulates MCS agents to actively pursue higher profits through continuous quality enhancement.

REFERENCES


Cong Zhao received his PhD in Computer Science and Technology from Xi’an Jiaotong University (XJTU) in 2017. He is currently a Research Associate in the Department of Computing at Imperial College London. His research interests include edge computing, computing economics, and people-centric sensing.

Shusen Yang received his PhD in Computing from Imperial College London in Dec. 2013. He is currently a Professor in the Institute of Information System Science and Technology at Xi’an Jiaotong University (XJTU). Before joining XJTU, Shusen worked as a Lecturer (Assistant Professor) at University of Liverpool from 2015 to 2016, and a Research Associate at Imperial College and Intel Collaborative Research Institute (ICRI) for sustainable connected cities from 2013 to 2015. Shusen is a recipient of China 1000 Young Talents Program, a DAMO Academy Young Fellow, and an honorary research fellow at Imperial College London. He is a senior member of IEEE and a member of ACM and CCF. Shusen is the founder and director of the IoT-DATA Lab, a director at the National Engineering Laboratory for Big Data Analytics (NEL-BDA), and the deputy head of Ministry of Education (MoE) Key Lab for Intelligent Networks and Network Security.

Julie A. McCann is a Professor in Computer Science at Imperial College London. Her research centers on highly decentralized and self-organizing scalable algorithms for spatial computing systems e.g. wireless sensing networks. She leads both the Adaptive Embedded Systems Engineering Research Group and the Intel Collaborative Research Institute for Sustainable Cities, and is currently working with NEC and others on substantive smart city projects. She has received significant funding through bodies such as the UK’s EPSRC, TSB and NERC as well as various international funds, and is an elected peer for the EPSRC. She has actively served on, and chaired, many conference committees and is currently Associate Editor for the ACM Transactions on Autonomous and Adaptive Systems. She is a Fellow of the BCS.