Privacy-preserving Truth Discovery with Outlier Detection in Mobile Crowdsensing Systems

Jingchen Zhao*, Bin Zhu*, Jian Li*, Shaoxian Yuan*, Kaiping Xue*†‡, Xianchao Zhang†‡
* School of Cyber Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China
† Key Laboratory of Medical Electronics and Digital Health of Zhejiang Province and Engineering Research Center of IntelligentHuman Health Situation Awareness of Zhejiang Province, Jiaxing University, Jiaxing, Zhejiang 314001, China
‡Corresponding author, kpxue@ustc.edu.cn (K. Xue), zhangxianchao@zjxu.edu.cn (X. Zhang)

***Abstract*—Recently, there have been many discussions in mobile crowd-sensing about privacy-preserving truth discovery because of its ability to extract truthful information from noisy or biased sensory data without privacy breaches. However, in practical applications, users (referred to as workers) may report outliers due to device malfunction, malicious workers, etc. These outliers will dramatically impact the accuracy of the truth discovery result. Detecting outliers based on existing privacy preservation schemes will carry an intolerable overhead, dramatically reducing the system's availability. In this paper, we propose our privacy-preserving truth discovery scheme that can detect outliers. Specifically, we adopt an anonymous mechanism to achieve privacy preservation. Since the existing anonymous mechanisms require huge overhead and do not work correctly when some workers exit, they are difficult to be applied in mobile crowdsensing systems. We design a lightweight and robust anonymous mechanism based on the edge computing paradigm. In addition, we eliminate the impact of outliers through outlier detection to achieve robustness of truth discovery results. Finally, we demonstrate the security of our scheme through security analysis and the efficiency of our scheme in terms of computation and communication overhead through extensive experiments.**

***Index Terms*—mobile crowdsensing, truth discovery, privacy preserving, outlier detection**

I. Introduction

As mobile and portable devices are increasingly equipped with various sensors, mobile crowdsensing is receiving more and more attention as a cost-effective data collection and analysis paradigm. This paradigm leverages mobile users' sensors to observe targets and collect sensory data. However, due to poor hardware quality, background noise, sensor calibration errors, etc., the sensory data provided by users (referred to as workers) often lacks sufficient accuracy. In order to discover truth information(referred to as the inferred truths) about objects from sensory data provided by multiple workers, truth discovery has been proposed [1], [2]. It can extract inferred truths from a large amount of sensory data with noise or bias by estimating the reliability of different workers during the computation.

However, the sensory data usually carry private information about workers [3], such as human behavior and location information. Cloud platforms may abuse this sensitive information, damaging workers' reputations and even putting workers at risk. Therefore, privacy preservation is an essential issue in mobile crowdsensing systems. In addition to privacy issues, the sensory data reported by workers may contain outliers due to possible device malfunction, operational error, etc. Worse, malicious workers may deliberately report outliers to impact crowdsensing tasks for illicit financial gain.

Some privacy-preserving truth discovery schemes have been proposed to address the privacy issue, but it is difficult to detect outliers in these schemes. Some schemes adopt homomorphic encryption [4]–[7]. Since the sensory data is encrypted, detecting outliers in these schemes will carry intolerable overhead. Meanwhile, these encryption schemes have huge computation and communication overhead. In masking or differential privacy schemes [8]–[11], it is difficult to detect outliers because noise is added to each sensory data. However, leaving out outliers in the truth discovery process will dramatically impact the accuracy of the inferred truths.

Actually, in mobile crowdsensing systems, multiple workers participate in the same observation task. If the server cannot determine which worker the sensory data belongs to, it cannot violate the worker's privacy, even if the server gets the plaintext sensory data. Therefore, in many scenarios (e.g., road condition detection, temperature detection, etc.), anonymous mechanisms can be employed to preserve workers' privacy.

The scheme [12] employs anonymous mechanisms [13] to preserve workers' privacy in privacy-preserving truth discovery. However, when some workers exited (dropped out or actively exited) from the mobile crowdsensing systems, it cannot correctly recover other workers' sensory data. Meanwhile, since the length of the data vector computed and transmitted by each worker is related to the whole number of workers in the system, [12] requires large computation and communication overhead on the workers and servers. So it cannot be applied in mobile crowdsensing systems where the workers are resource-constrained and may exit at any time. To the best of our knowledge, in mobile crowdsensing systems, there is no privacy-preserving truth discovery scheme that can eliminate the impact of outliers.

To address the above challenge, we leverage the computing power of cloud-edge architecture and construct a lightweight privacy-preserving truth discovery scheme that is unimpacted by worker exits and capable of detecting outliers. Specifically, our contributions are summarized as follows:

- We propose a privacy-preserving truth discovery scheme

978-1-6654-3540-6/22/$31.00 © 2022 IEEE

with robustness to outliers by adopting an anonymous mechanism. In this way, the server can detect outliers based on plaintext sensory data without requiring a huge computational overhead.

- By designing an anonymous mechanism based on cloud-edge architecture, we achieve more lightweight on the workers and server-side. Meanwhile, our design is also very robust against worker exits.
- We prove that our scheme is secure through security analysis. Meanwhile, Through extensive experiments, we prove that our scheme is efficient regarding computational and communication overheads and robustness.

The remainder of this paper is organized as follows. Section II introduces the related work. Section III presents our models and design goals. In Section IV, we introduce some preliminaries. After that, we describe our scheme in Section V. We analyze its security in Section VI and evaluate its performance in Section VII. Finally, we conclude this work in Section VIII.

II. Related Work

Recently, truth discovery [1], [2] has received considerable attention from researchers due to its efficiency in extracting truthful information from large amounts of noisy or biased sensory data. However, in mobile crowdsensing systems, workers submit sensory data that usually carry private information about the workers.

Researchers have proposed many truth discovery schemes to preserve privacy with concerns about privacy breaches, where [4]–[7] achieve privacy-preserving by adopting homomorphic encryption. Miao *et al.* [4] first proposed a privacy-preserving single-server truth discovery system, but their scheme requires workers to be involved in numerous computations and communication processes. Zhang *et al.* [7] adopted a cloud-edge architecture to improve the scalability of the privacy-preserving truth discovery system. The schemes of [5], [6] shift the computation overhead from the workers to the server-side by adopting two non-colluding servers. However, all of these homomorphic encryption schemes require huge computational overhead. Some researchers have adopted masking to preserve the privacy of workers. The schemes [8], [9] avoid the huge computational overhead of adopting homomorphic encryption through multiple rounds of communication between worker and server. However, workers are free to modify their weights, which causes these schemes to work correctly only when workers are not evil. Some schemes [10], [11] adopt differential privacy to reduce workers' and server-side's overhead while preserving workers' privacy. However, these schemes require adding noise to workers' sensory data to preserve privacy, which inevitably reduces the accuracy of the inferred truths. In summary, these schemes mentioned above can preserve the privacy of workers, but they fall short of efficiency or accuracy, and it is difficult to detect outliers on these schemes due to encryption or added noises.

Anonymous mechanisms [13] are also a common approach to preserving workers' privacy. Since the anonymous mechanisms delink workers' sensory data from their identity, it allows the worker's plaintext sensory data to be available while preserving the worker's privacy. However, in anonymous mechanisms, a trustworthy centralized data publisher is usually required to collect the workers' sensory data and then anonymize it, which is often difficult to be satisfied in mobile crowdsensing systems. Zhang *et al.* [14] proposed a scheme to delink worker data from worker identity, which does not require a centralized data publisher. Tang *et al.* [12] established a privacy-preserving truth discovery scheme by adopting an anonymous mechanism inspired by [14]. However, the scheme [12] cannot correctly recover the sensory data of other workers when there are some workers exited. Meanwhile, the workers and server-side have a large computation and communication overhead. This is because each worker's length of the data vector computed and transmitted is related to the whole number of workers in the system. Therefore, it is not suitable for mobile crowdsensing systems. To the best of our knowledge, there is no existing privacy-preserving truth discovery scheme that can eliminate the impact of outliers in mobile crowdsensing systems.

III. Problem Statement

A. System Model

The entities in our system include many workers, some edge nodes (ENs), and a cloud server (CS), and the relationship between them is shown in Fig. 1.

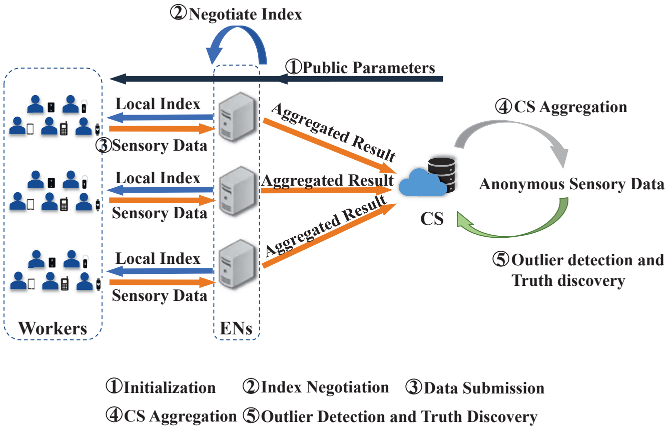


Fig. 1. System Model

- Workers: Workers are data providers who sense objects and report sensory data to EN through their devices. These devices are usually resource-constrained in terms of computation and communication. In addition, workers may exit from the system at any time.
- Edge Node (EN): EN aggregates the sensory data reported by workers and uploads the results to CS. Meanwhile, ENs negotiate the index location, which is used to anonymize the workers' sensory data.
- Cloud Server (CS): CS aggregates the results uploaded by ENs to obtain anonymized workers' sensory data and

performs outlier detection and truth discovery to get the inferred truths.

In this work, the inferred truth for the object m is referred to as x_m^*. Suppose there are M objects (denoted $\{o_1, o_2, ...o_M\}$) and K workers (denoted $\{u_1, u_2, ...u_K\}$) in the system, the sensory data of worker u_k for object o_m is referred to as x_m^k. Besides, we use $E = \{E_1, E_2...E_n\}$ and G_{E_i} to denote the ENs and the number of workers maintained by E_i, respectively. G_{E_i} can be different for each EN.

B. Threat Model

In mobile crowdsensing systems, similar to previous works [7], [11], we assume ENs and CS will perform the protocol honestly, but they will also be curious about workers' privacy. Workers' privacy is defined as their sensory data and weights can only be accessed anonymously. Meanwhile, there is no collusion between ENs and CS. In addition, outliers may appear in the sensory data reported by workers due to device errors, worker operation errors, malicious workers, etc. Each entity in the system is incapable of breaking the Decisional Diffie-Hellman (DDH) assumption [15] and performing ciphertext-only attacks (COA).

C. Design Goals

- Privacy-Preservation: The proposed scheme should ensure that workers' weights and sensory data are not leaked to ENs and other workers and that all workers' data is anonymous to CS. In addition, the inferred truths should not be revealed to ENs and workers.
- Robust: The proposed scheme should be able to detect outliers and eliminate the impact of outliers on the accuracy of the inferred truths. In addition, workers' exit from the crowdsensing system at any time should not impact the outlier detection and truth discovery process.
- Efficiency: The proposed scheme should be efficient regarding computational and communication overhead on the worker and server-side.

IV. Preliminaries

A. Truth Discovery

Truth discovery obtains the inferred truths by iteratively carrying the weight update step and the truth update step, the details of these two steps are as follows:

1) Weight Update: In this step, given the currently inferred truths $\{x_m^*\}_{m=1}^M$ and the workers' sensory data $\{x_m^k\}_{k=1}^{k=K}$. The weight of worker k is computed as

$$w_k = \log(\frac{\sum_{k=1}^K \sum_{m=1}^M d(x_m^k, x_m^*)}{\sum_{m=1}^M d(x_m^k, x_m^*)}),$$

in this paper, we adopt the distance function by the squared distance $d(x_m^k, x_m^*) = (x_m^k - x_m^*)^2$.

2) Truth Update: After the weight update step, each inferred truth is computed as

$$x_m^* = \frac{\sum_{k=1}^K x_m^k w_k}{\sum_{k=1}^K w_k}.$$

Truth discovery iterations perform the above two-step until it satisfies the convergence condition, which is usually a predefined number of iterations or the change of inferred truths between two iterations. Since there is no need to encrypt the sensory data, our scheme uses the same method for fractions and integers without loss of precision.

B. Key agreement

Key agreement is a common method of establishing shared secrets between two parties. Secrets shared between parties can be used for secure communication or as random seeds for masking. In this paper, we adopt the Diffie-Hellman key agreement [15] to establish a shared secret. The key agreement consists of the following three algorithms:

- $\mathbf{KA}.\mathbf{param}(1^\lambda) \mapsto (\mathbb{G}, g, p)$: Generate a group $\mathbb{G}$ with prime order q and a generator g by security parameter λ.
- $\mathbf{KA}.\mathbf{gen}(\mathbb{G}, g, p) \mapsto (SK_i, PK_i)$: Generate a public-private key pair by taking $\mathbb{G}, g, p$ as input.
- $\mathbf{KA}.\mathbf{agree}(SK_i, PK_j) \mapsto S_{ij}$: Given the private key SK_i of client i and the public key PK_j of client j to get a shared secret S_{ij}.

V. The Proposed Scheme

A. Overview

As discussed above, we aim to detect outliers and improve the efficiency of privacy-preserving truth discovery in mobile crowdsensing systems.

Our system includes many workers, some ENs, and a CS. The system flow is shown in Fig. 1. Initially, each entity is initialized with the public parameters published by CS. During each truth discovery task, the ENs negotiate indexes by random permutation function and then assign a local index to each worker they maintain. Each worker adds random numbers to their sensory data and anonymizes it according to the received index, then reports the result to EN. Workers can remain offline after completing the reporting process until they are involved in the next truth discovery. Each EN preliminarily aggregates the results reported by the workers. After that, each EN masks the aggregated results and sends them to CS. CS performs further aggregation and eliminates the random numbers of workers to obtain anonymized data for all workers. Finally, CS removes outliers and performs truth discovery to get the inferred truths.

The complete scheme can be divided into five specific phases: *Initialization, Index Negotiation, Data Submission, CS Aggregation, Outliers Detection and Truth Discovery*.

B. Initialization Phase

In this phase, CS generates $(\mathbb{G}, g, p)$ via security parameter λ and $\mathbf{KA}.\mathbf{param}(\cdot)$, where $\mathbb{G}$ is a cyclic group of prime order p, and g is a generator of $\mathbb{G}$. Then, CS generates a key pair (SK_{CP}, PK_{CP}) via $\mathbf{KA}.\mathbf{gen}(\cdot)$. After that, CS publishes PK_{CP} and $pp = (\mathbb{G}, g, p, \boldsymbol{\pi}, \mathbf{F}, \mathbf{AE})$, where $\boldsymbol{\pi}$ is a random permutation function, $\mathbf{F}$ is a pseudo-random function, and $\mathbf{AE}$ is a symmetric encryption algorithm. Each EN and worker generates a key pair via $\mathbf{KA}.\mathbf{gen}(\cdot)$. ENs establish shared secrets S_{E_i,E_j} via $\mathbf{KA}.\mathbf{agree}(\cdot)$, each worker reports their public key to CS through the ENs and establishes shared secret S_{CS,u_k} with CS. After this phase, workers can participate at any time, they only need to choose an EN to be joined and establish a shared secret with CS.

C. Index Negotiation Phase

In this phase, ENs negotiate the index, which will be used to anonymize workers' sensory data.

Step1: Each EN generates a local sequence $\{1, 2, ..., K\}$ and shuffles it by $\boldsymbol{\pi}$ to obtain $\boldsymbol{L}$. Then, EN encrypts $\boldsymbol{L}$ using the shared secrets with other ENs to obtain $\widetilde{\boldsymbol{L}}$ and sends it to CS. Take E_i as an example, E_i generates a shuffled local sequence $\boldsymbol{L}$, suppose E_j maintains G_{E_j} workers, $G_{E_j} = b-a+1$, E_{j+1} maintains $G_{E_{j+1}}$ workers, $G_{E_{j+1}} = d - c + 1$. E_i encrypts the elements from a to b of $\boldsymbol{L}$ using the shared secret S_{E_i,E_j}, encrypts the elements from c to d of $\boldsymbol{L}$ using the shared secret $S_{E_i,E_{j+1}}$, and so on to get the following encryption result $\widetilde{\boldsymbol{L}}$.

$$\begin{aligned}\widetilde{\boldsymbol{L}} = \{&\mathbf{AE}.\mathbf{encrypt}(S_{E_i,E_j}, \boldsymbol{L}_a, ..., \boldsymbol{L}_b)\\ &\mathbf{AE}.\mathbf{encrypt}(S_{E_i,E_{j+1}}, \boldsymbol{L}_c, ..., \boldsymbol{L}_d)\\ &...\}.\end{aligned}$$

If $i = j$, E_i randomly selects a random number for encryption. Finally, each EN uploads its $\widetilde{\boldsymbol{L}}$ to CS.

Step2: CS randomly selects one from all $\widetilde{\boldsymbol{L}}$ and broadcasts it to every EN. After each EN gets $\widetilde{\boldsymbol{L}}$, it can decrypt G_{E_i} indexes using the shared secret established with the EN selected by CS, and these indexes form a vector $\boldsymbol{l}$.

Step3: Each EN generates a sequence $\boldsymbol{l'} = \{1, 2, ...G_{E_i}\}$, shuffle it by $\boldsymbol{\pi}$, and randomly assigns an index $\boldsymbol{l'}_{u_k}$ from it for each worker it maintains.

D. Data Submission Phase

In this phase, each worker adds random numbers to their sensory data and anonymizes them. After that, each worker reports the results to his/her EN. EN performs a preliminary aggregation and uploads the aggregated results to CS.

Step1: Each worker u_k generates a random number r_{u_k} and computes $\boldsymbol{w}$, where

$$\boldsymbol{w}_i = \begin{cases} \mathbf{F}(r_{u_k}||i||m) & i \neq \boldsymbol{l'}_{u_k}, \\ \mathbf{F}(r_{u_k}||i||m) + x_k^m & i = \boldsymbol{l'}_{u_k}. \end{cases}$$

Then, u_k reports $\boldsymbol{w}$ and $\mathbf{AE}.\mathbf{encrypt}(S_{CS,u_k}, r_{u_k})$ to EN. When u_k completes the reporting process, u_k can go offline. In this step, even if some workers do not report their data due to exit from the system, it will not impact the final outlier detection and truth discovery process.

Step2: EN receives $\boldsymbol{w}$ and $\mathbf{AE}.\mathbf{encrypt}(S_{CS,u_k}, r_{u_k})$ reported by all the workers it maintains and performs a column-by-column summation of all received $\boldsymbol{w}$ to obtain $\boldsymbol{w'}$, where each element

$$\boldsymbol{w'}_i = \ \textstyle\sum_{u_k \in E_n} \mathbf{F}(r_{u_k}||i||m) \ + x_s^m,$$

where x_s^m is the sensory data from one of the workers. After that, EN forwards all $\mathbf{AE}.\mathbf{encrypt}(S_{CS,u_k}, r_{u_k})$ to CS.

Step3: CS decrypts $\mathbf{AE}.\mathbf{encrypt}(S_{CS,u_k}, r_{u_k})$ to get r_{u_k} and generates M random vectors $\boldsymbol{r'}$ and computes $\boldsymbol{b'}$ for each E_i based on $\boldsymbol{r'}$ and r_{u_k}. Finally, CS returns $\boldsymbol{b'}$ to E_i, where

$$\boldsymbol{b'}_i = \textstyle\sum_{u_k \in E_n} \mathbf{F}(r_{u_k}||i||m) - \boldsymbol{r'}_i.$$

Step4: Each EN uses shared secrets S_{E_i,E_j} among them to generate the following random vectors $\boldsymbol{Z}$ and $\boldsymbol{Z'}$, where

$$\begin{aligned}\boldsymbol{Z}_k \ &= \sum_{i>j} \mathbf{F}(S_{E_i,E_j}||k) - \sum_{i<j} \mathbf{F}(S_{E_i,E_j}||k),\\ \boldsymbol{Z'}_k &= \sum_{i>j} \mathbf{F}(S_{E_i,E_j}||k||m) - \sum_{i<j} \mathbf{F}(S_{E_i,E_j}||k||m).\end{aligned}$$

We can find that the sum of all EN-generated random numbers is zero, where $\sum_{E_i} \boldsymbol{Z}_k = 0$ and $\sum_{E_i} \boldsymbol{Z'}_k = 0$.

Step5: Each EN generates the following random number index vector $\boldsymbol{I}$ and the following data vector $\boldsymbol{D}$, where

$$\boldsymbol{I}_k = \begin{cases} \boldsymbol{Z}_k & k \notin \{\boldsymbol{l}_i | i \in (0, G_{E_i})\}, \\ \boldsymbol{Z}_k + i & k = \boldsymbol{l}_i, \end{cases}$$

$$\boldsymbol{D}_k = \begin{cases} \boldsymbol{Z'}_k & k \notin \{\boldsymbol{l}_i | i \in (0, G_{E_i})\}, \\ \boldsymbol{Z'}_k + \boldsymbol{w'}_i - \boldsymbol{b'}_i & k = \boldsymbol{l}_i. \end{cases}$$

After that, each EN sends $\boldsymbol{I}$ and $\boldsymbol{D}$ to CS.

E. CS Aggregation Phase

CS receives the vectors $\boldsymbol{I}$ and $\boldsymbol{D}$ uploaded by each EN, and aggregates them column-by-column to obtain $\boldsymbol{I'}$ and $\boldsymbol{D'}$, respectively, where

$$\begin{aligned}&\boldsymbol{I'}_k = i, \qquad if : k = \boldsymbol{l}_i,\\ &\boldsymbol{D'}_k = x_s^m + \boldsymbol{r'}_{\boldsymbol{I'}_k}.\end{aligned}$$

After that, CS eliminates $\boldsymbol{r'}_{\boldsymbol{I'}_k}$ from $\boldsymbol{D'}_k$ to obtain the anonymized sensory data $\{x_s^m | s \in (0, K)\}$.

F. Outlier Detection and Truth Discovery Phase

In this phase, CS detects outliers and performs truth discovery process.

For each task $\{o_1, o_2, ...o_M\}$, CS first sorts the workers' sensory data in ascending order to obtain $x_1^m, x_2^m, ...x_K^m$, quadrates $x_1^m, x_2^m, ...x_K^m$, and the value of the three split points are represented as Q_1,Q_2,Q_3. CS removes all sensory data that are beyond the following detection interval

$$(Q_1 - \alpha(Q_3 - Q_1), Q_3 + \alpha(Q_3 - Q_1)),$$

where α indicates the intensity of outlier detection. After detecting outliers and removing all outliers from the anonymized sensory data, CS performs truth discovery to obtain the inferred truths.

VI. Security Analysis

Theorem 1. *Suppose these entities satisfy our threat model, our scheme will preserve the privacy of workers' sensory data and weights, and the privacy of inferred truths.*

Proof. We first prove that the privacy of workers and the inferred truths will not be disclosed to ENs and workers. Then, we prove that CS cannot violate the privacy of workers.

(1) Since workers are only participate in the initialization phase and report sensory data to EN, they cannot obtain other workers' sensitive information. Since EN cannot obtain $\boldsymbol{r}'$ from CS, EN cannot obtain x_k^m through $\boldsymbol{b}'$ and $\boldsymbol{w}'$. Even if all ENs collude, they also cannot get $\boldsymbol{r}'$ from the respectively received $\boldsymbol{b}'$. Meanwhile, if EN wants to obtain x_k^m by $\boldsymbol{w}$, EN needs to get r_{u_k}, which requires EN to break the DDH assumption [15] or perform COA, so the workers' sensory data x_k^m doesn't disclose to ENs. We consider the possibility of EN colluding with several workers, Since each worker randomly generates his/her r_{u_k} and adds $\mathbf{F}(r_{u_k}||i||m)$ to his/her sensory data, EN cannot access other non-collusion workers' sensory data. Since the truth discovery is performed by CS, the workers' weights and inferred truths are not leaked to workers or ENs.

(2) If CS wants to determine each sensory data in $\boldsymbol{D}'$ from which worker or each element in $\boldsymbol{I}'$ from which EN, CS must obtain $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ or obtain $\boldsymbol{L}$. CS can only compute $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ by S_{E_i,E_j} or obtain $\boldsymbol{L}$ by decrypting $\widetilde{\boldsymbol{L}}$, which requires the ability to break the DDH assumption [15] or perform COA. Furthermore, We suppose exists a stimulator $\mathcal{A}$, which randomly selects a random permutation function $\boldsymbol{\pi}'$ and calculates $\boldsymbol{L}'$ ($\boldsymbol{D}$ and $\boldsymbol{I}$ are the same as $\boldsymbol{L}$). For the view of CS, there is computational indistinguishability $VIEW_{CS}(\boldsymbol{L}) \overset{c}{\equiv} VIEW_{CS}(\boldsymbol{L}')$. Otherwise, if CS can distinguish workers' anonymized data, it is equivalent to random permutation function $\boldsymbol{\pi}$ without randomness. It contradicts our assumption. So the workers' sensory data and weights are anonymous (anonymous set size is K.) to CS, which cannot invade the privacy of workers. Finally, since each worker can only get his/her index within EN, collusion between CS and workers does not invade the privacy of other workers.

In summary, the privacy of inferred truths, workers' sensory data, and workers' weights can be preserved.

VII. Performance Evaluation

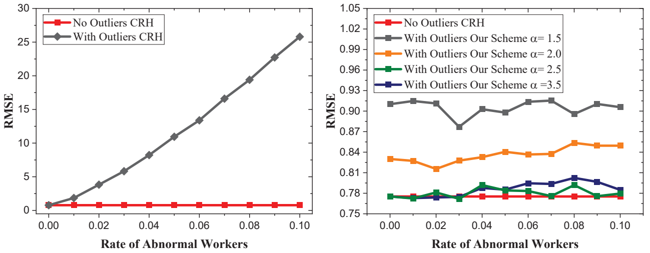
In this section, we will measure our scheme in terms of accuracy, computation overhead, and communication overhead, respectively. The configuration is a computer with 2.90Ghz Intel i5 and 16GB RAM and Windows 10 with Python 3.9.6. We adopt the python library of random for pseudo-random function $\mathbf{F}$ and random permutation function $\boldsymbol{\pi}$, and the library of cryptography [16] for key agreement protocol.

Since [5] notes that sensory data from various normal workers are likely to be distributed normally. In our experiments, we generate sensory data of normal workers from the normal distribution. We represent sensor errors, worker operation errors or malicious workers by generating sensory data above the normal distribution. These workers are called abnormal workers. We set $M = 20$, the number of ENs is 10, the size of p is set as 512 bits, and the size of symmetric keys and plaintexts are set as 256 bits and 64 bits, respectively.

A. Accuracy

Similar to AnonymTD [12], we use the Root of Mean Squared Error (RMSE) to evaluate the accuracy of the inferred truths $\{x_m^*\}_{m=1}^M$. RMSE is defined as RMSE $= (\sum_{m=1}^M (x_m^* - \hat{x}_m)^2/M)^{1/2}$, where $\{\hat{x}_m\}_{m=1}^M$ is the real truth of objects. We set $K = 1000$ and the no outliers CRH [1] was taken as a baseline scheme. Our scheme can be easily adapted to other truth discovery approaches as well.

Outlier Impact. We first measure the impact of outliers on the accuracy of the inferred truths. The rate of abnormal workers varies from 0% to 10%, and no worker exits. As shown in Fig. 2a, we observed that the outliers reported by abnormal workers can dramatically impact the accuracy of the inferred truths. After that, We set different α to evaluate our scheme's effectiveness in detecting and eliminating the impact of outliers. In Fig. 2b, we observe that the impact of outliers can be effectively eliminated by outlier detection. There is a decrease in accuracy when α is small, which is due to the overpowering outlier detection causing some normal workers' sensory data is removed.

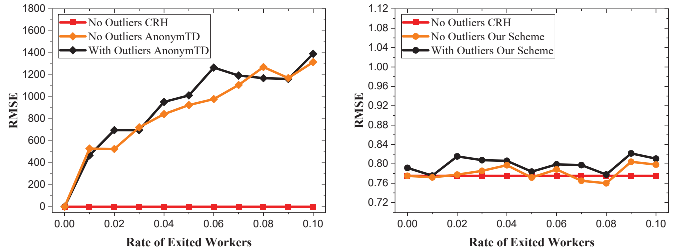


(a) Without Outlier Detection

(b) With Outlier Detection

Fig. 2. Impact of Outliers

Worker exits Impact. We measure the performance of our scheme and AnonymTD when some workers exit. We set $\alpha = 2.5$, and the rate of exited workers varies from 0% to 10%. Fig. 3a shows that AnonymTD cannot work correctly when some workers exit. Because in AnonymTD, the server cannot eliminate the noise added by the exit of workers, the RMSE increases with the number of exited workers. Fig. 3b shows that our scheme is almost unimpacted by worker exits. When worker exits and outliers coexist, our scheme remains effective in eliminating their impact. These experimental results show that our scheme has stronger robustness.



(a) Impact on AnonymTD

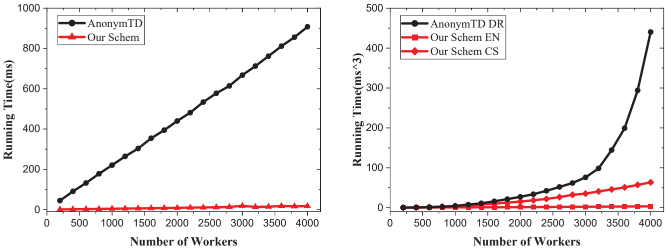
(b) Impact on Our Scheme

Fig. 3. Impact of Worker Exits

B. Efficiency Evaluation

In this part, we measure our scheme in terms of computation and communication overhead. We set the number of workers to vary from 0 to 4000.

Computation Overhead. As shown in Fig. 4a, since the number of pseudo-random numbers to be computed by each worker in our scheme is much smaller than AnonymTD, the workers' computation overhead in our scheme is much less than AnonymTD. As shown in Fig. 4b, since the preliminary aggregation of ENs, our scheme is also more efficient on the server-side. Meanwhile, with the number of workers increasing, our scheme's workers and server-side overhead increase much less than AnonymTD.



(a) Running Time on Workers (b) Running Time on server-side

Fig. 4. Computation Overhead

Communication Overhead. We measure the communication overhead of our scheme and AnonymTD for different numbers of workers and objects. As shown in Table I. We can see that our scheme is more efficient than AnonymTD in terms of both worker and server-side communication overhead. This is because in our scheme, the length of the data vector to be transmitted by each worker is much smaller than in AnonymTD, and the server only needs to receive the results after the preliminary aggregation of the edge nodes. Although we adopt ENs, the communication overhead of all ENs and CS is still much smaller than the server in AnonymTD.

Experimental results show that our scheme is more efficient regarding computational and communication overheads on both the workers and server-side.

TABLE I
COMMUNICATION OVERHEAD (KB)

Number of Workers and Objects		Workers		EN		Cloud Server	
		Our Scheme	AnonymTD	Our Scheme	AnonymTD	Our Scheme	AnonymTD
M = 10	K = 100	**0.828**	**7.835**	$\mathbf{0.02\times10^3}$	-	$\mathbf{0.18\times10^3}$	$\mathbf{0.78\times10^3}$
	K = 500	**3.953**	**39.085**	$\mathbf{0.28\times10^3}$	-	$\mathbf{0.89\times10^3}$	$\mathbf{1.95\times10^4}$
	K = 2000	**15.671**	**156.273**	$\mathbf{3.49\times10^3}$	-	$\mathbf{3.59\times10^3}$	$\mathbf{3.12\times10^5}$
M = 20	K = 100	**1.609**	**15.648**	$\mathbf{0.05\times10^3}$	-	$\mathbf{0.34\times10^3}$	$\mathbf{1.56\times10^3}$
	K = 500	**7.859**	**78.148**	$\mathbf{0.56\times10^3}$	-	$\mathbf{1.71\times10^3}$	$\mathbf{3.90\times10^4}$
	K = 2000	**31.296**	**312.523**	$\mathbf{6.94\times10^3}$	-	$\mathbf{6.87\times10^3}$	$\mathbf{6.25\times10^5}$

VIII. CONCLUSION

In this paper, we proposed a lightweight and robust privacy-preserving truth discovery with an outliers detection scheme. Firstly, we carefully discussed the impact of outliers on the accuracy of the inferred truths. After that, we proposed a scheme that can eliminate the impact of outliers and worker exits by adopting the cloud-edge architecture and anonymous mechanism. Finally, through the security analysis, we proved that our solution is secure. Extensive experiments show that our proposed scheme is lightweight regarding computation and communication overhead on the workers and server-side. Meanwhile, our scheme can effectively eliminate the impact of outliers and worker exits on the accuracy of the inferred truths. These features make our scheme more feasible in mobile crowdsensing systems.

ACKNOWLEDGMENT

The work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61972371, Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) under Grant No. Y202093, and Key Laboratory of Medical Electronics and Digital Health of Zhejiang Province under grant No. MEDH202201.

REFERENCES

[1] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (PODS)*, pp. 1187–1198, ACM, 2014.

[2] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *Proceedings of the VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.

[3] Z. Wang, X. Pang, J. Hu, W. Liu, Q. Wang, Y. Li, and H. Chen, "When mobile crowdsensing meets privacy," *IEEE Communications Magazine*, vol. 57, no. 9, pp. 72–78, 2019.

[4] C. Miao, W. Jiang, *et al.*, "Cloud-enabled privacy-preserving truth discovery in crowd sensing systems," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pp. 183–196, ACM, 2015.

[5] Y. Zheng, H. Duan, X. Yuan, and C. Wang, "Privacy-aware and efficient mobile crowdsensing with truth discovery," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 1, pp. 121–133, 2020.

[6] K. Xue, B. Zhu, *et al.*, "InPPTD: A lightweight incentive-based privacy-preserving truth discovery for crowdsensing systems," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4305–4316, 2021.

[7] C. Zhang, L. Zhu, C. Xu, X. Liu, and K. Sharif, "Reliable and privacy-preserving truth discovery for mobile crowdsensing systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1245–1260, 2019.

[8] G. Xu, H. Li, S. Liu, M. Wen, and R. Lu, "Efficient and privacy-preserving truth discovery in mobile crowd sensing systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3854–3865, 2019.

[9] Y. Liu, S. Tang, H.-T. Wu, and X. Zhang, "RTPT: A framework for real-time privacy-preserving truth discovery on crowdsensed data streams," *Computer Networks*, vol. 148, pp. 349–360, 2019.

[10] P. Sun, Z. Wang, L. Wu, Y. Feng, X. Pang, H. Qi, and Z. Wang, "Towards personalized privacy-preserving incentive for truth discovery in mobile crowdsensing systems," *IEEE Transactions on Mobile Computing*, vol. 21, no. 1, pp. 352–365, 2020.

[11] D. Wang, J. Ren, Z. Wang, X. Pang, Y. Zhang, and X. S. Shen, "Privacy-preserving streaming truth discovery in crowdsourcing with differential privacy," *IEEE Transactions on Mobile Computing*, 2021. DOI: 10.1109/TMC.2021.3062775.

[12] J. Tang, S. Fu, *et al.*, "Achieving privacy-preserving and lightweight truth discovery in mobile crowdsensing," *IEEE Transactions on Knowledge and Data Engineering*, 2021. DOI: 10.1109/TKDE.2021.3054409.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 3–es, 2007.

[14] Y. Zhang, Q. Chen, and S. Zhong, "Privacy-preserving data aggregation in mobile phone sensing," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 980–992, 2016.

[15] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. 22, no. 6, pp. 644–654, 1976.

[16] "Cryptography." https://cryptography.io/en/latest/fernet/. Accessed on May., 2022.