An Incentive-Based Differential Privacy-Preserving Truth Discovery over Streaming Data

Yaxuan Huang*, Feng Liu*, Jingcheng Zhao*, 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*—Truth discovery is an effective tool to infer true information from multi-source data and has been widely applied in mobile crowdsensing systems. In some specific scenarios, the sensory data are collected in a streaming fashion with time-varying information, and the server should update the truth in time. Under such circumstances, local differential privacy-based mechanism can satisfy the requirement of real-time processing properly while keeping the privacy of sensory data. However, directly applying local differential privacy to handle streaming data will disclose the long-term potential privacy and decrease the accuracy. To address these problems, we propose an incentive-based privacy-preserving truth discovery framework over streaming data. Firstly, we adopt the sequential composition theorem of w-event privacy to protect workers' long-term privacy. Second, we design an incentive mechanism to improve the submitted data utility and thus avoid the decrease in accuracy. In this way, our scheme ensures that workers submit more accurate data while their global privacy is still guaranteed. Finally, we prove our scheme satisfies w-event (ϵ, δ) differential privacy and theoretically analyze the result utility. Extensive experiments also demonstrate the effectiveness of our incentive mechanism.**

***Index Terms*—Truth Discovery, Privacy Preservation, Local Differential Privacy**

I. Introduction

With the popularity of mobile and wearable devices, crowdsensing systems are developing rapidly. In a typical crowdsensing application, workers upload the sensory data to the server, and the server analyzes these data for further use. However, due to the ambient noise and the different performance of devices, the quality of sensory data varies from worker to worker, thus the result of simple aggregation can deviate from the ground truth. For this reason, truth discovery algorithms [1] are proposed to infer reliable aggregated information from multi-source data by assigning different weights to workers. Although the truth discovery algorithms can provide true answers for crowdsensing systems, they cause privacy concerns since the data submitted by workers may contain some sensitive information. In order to preserve the privacy of workers, some works [2]–[4] suggest using cryptographic tools to provide strong protection for individual privacy.

In some truth discovery scenarios, the sensory data are collected in a streaming fashion with time-varying information, such as temperature, humidity and traffic information. As for the privacy-preserving truth discovery (PPTD) over streaming data, workers submit data at any time and servers should update the truth in time, which requires high computational efficiency with privacy-preserving. Unfortunately, existing cryptographic methods involve time-consuming calculation or additional communication costs, they are not suitable for streaming data. In this respect, local differential privacy (LDP)-based approaches are more suitable compared with the traditional cryptography-based PPTD schemes. Recently, some LDP-based PPTD schemes [5]–[9] are proposed to protect privacy by adding noise to the original data.

In scenarios of truth discovery over streaming data, two important issues need to be considered in LDP-based schemes. One issue is long-term potential privacy disclosure. In truth discovery over streaming data, data collection is often a long-term procedure compared with one-time truth discovery, so we need to take the privacy of time dimension into consideration. Specifically, the noise continuously added to data over time is considered to be relevant, and disturbed data may also have potential security problem through statistical analysis. To preserve the privacy of streaming data in truth discovery, we use the model of w-event privacy [10], [11], which guarantees provable privacy for any event sequence occurring at the time window of size w.

The other issue is the decrease of accuracy, because the LDP mechanisms introduce additional noise to data. Generally, workers like to add more noise to their data for stronger privacy protection. Although the error caused by noise can be reduced during aggregation of large amounts of data, with the number of participants increasing, the effect of error reduction is limited (cf., Section VI). Therefore, we design an incentive mechanism to improve the accuracy of results by motivating workers to submit data with high utility. Encouraged by monetary rewards, workers perturb sensory data with less noise under the condition of guarantying the global privacy for consecutive periods.

To handle these issues, this paper makes the following contributions:

- We propose a privacy-preserving truth discovery scheme over streaming data for consecutive periods. The model of w-event differential privacy provides provable privacy guarantee for any event sequence occurring at the time window of size w. And workers' privacy is not disclosed to any other participants.

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

- A lightweight incentive mechanism is designed for large-scale crowdsensing truth discovery tasks based on sequential composition property. Motivated by monetary rewards, workers submit less but higher-utility data while their global privacy is still guaranteed.
- We theoretically analyze the data utility and prove our scheme satisfies w-event (ϵ, δ)-LDP. Moreover, through extensive experiments, we demonstrate that our incentive mechanism can improve the accuracy of truth discovery results.

The remainder of the paper is organized as follows. Section II introduces the related work, including existing works of PPTD and incentive mechanisms of it. Then, the problem statement is given in Section III. In Section IV, we describe the truth discovery algorithm and local differential privacy definition. We present details of incentive-based differential PPTD over streaming data in Section V. After that, experimental results and theoretical analysis are provided in Section VI. Finally, Section VII concludes this paper.

II. Related Work

We classify PPTD schemes into three kinds according to privacy-preserving methods: cryptography-based PPTD [2]–[4], data masking-based PPTD [12] and LDP-based PPTD [5]–[9]. Compared with the other two kinds, LDP approaches have the advantages of low computational cost and low communication consumption in truth discovery. In these LDP-based PPTD schemes, servers compute the truth over perturbed data. Although satisfying differential privacy, these studies provide provable privacy protection, they are designed for one-time truth discovery [5]–[8] or do not support protection for workers' privacy during consecutive periods [9].

Motivated by the problem of long-term privacy protection, some works [10], [11] are proposed to protect any event sequence occurring in successive time instants. They provide the model of w-event differential privacy for data publishing. And inspired by these works, [13] considers the correlations among truths over time and the characteristic of workers' reliabilities, it achieves high accuracy of truth discovery over streaming data. However, controlling privacy budget of workers, edge servers assumed to be semi-honest can get more privacy than the cloud server, which is not measured for privacy disclosure.

In order to reduce lazy workers or improve the accuracy of results, some reserchers designed incentive mechanisms for PPTD [5], [6], [14]. Specifically, [14] rewards workers according to the cumulative weight to reduce lazy workers in the system. By designing a set of contracts with workers, [5], [6] optimize the truth discovery accuracy under the given budget respectively over binary discrete and continuous data. Nevertheless, to the best of our knowledge, none of these works are suitable for scenarios of truth discovery over streaming data.

III. Problem Statement

A. System Model

As shown in Fig. 1, our framework contains two different types of parties: one cloud server and many workers of crowdsensing tasks. Among them, the cloud server is responsible for collecting workers' data and calculating the estimated truth. And workers use devices to sense data from tasks required objects, then they upload perturbed data to the server immediately out of the timeliness. For notational convenience, in the t-th time period, the sensory data of the i-th worker from objects is denoted as $\boldsymbol{x}_i^t$ which is perturbed as $\hat{\boldsymbol{x}}_i^t$, and the server totally receives N_t workers' data. Assume that there are M objects to be sensed, and the number of workers is not fixed, since new workers may join the task at any time, and participants may exit when their privacy budgets are consumed up.

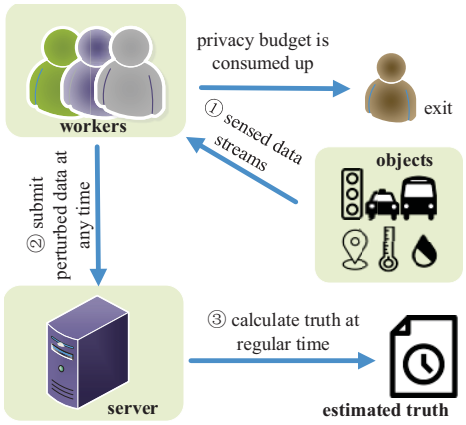

Fig. 1. System Model

B. Security Assumption

We assume the server is honest-but-curious, in other words, it honestly executes the protocol, but also tries to infer private information from other participants. We also consider workers to be lazy and malicious, which means they may submit fake data rather than costly sensory data and try to lie to the server about their privacy budget in order to get more rewards.

C. Design Goals

The main goal of our proposed scheme is to design a PPTD scheme over streaming data. Specifically, our design goals are three aspects:

1) *Privacy Preservation*: Every worker's original sensory data are protected by the Gaussian mechanism and privacy should not be disclosed in the long-term collection process. And the proposed scheme should satisfy w-event (ϵ, δ)-differential privacy.
2) *Accuracy*: In the proposed scheme, the estimated truth should converge to the ground truth. The expectation of error of estimated truth at a single perturbed point can be measured with differential privacy parameters.

3) *Effectiveness of Incentive Mechanism*: According to the proposed incentive mechanism, on the one hand, for workers, the higher utility the data they can provide, the higher rewards they can get. On the other hand, for server, the more rewards server can provide, the more accurate result of truth discovery can be delivered.

IV. Preliminaries

A. Truth Discovery over Streaming Data

The truth discovery algorithm iCRH (incremental CRH) [1] is designed for data streams. Because data streams come quickly, server doesn't calculate iteratively as CRH [1] required, but measures workers' reliability in the long-term by their weight.

According to iCRH procotol, in the t-th time period, the server totally receives N_t workers' data and do the following three steps in order: truth update, distance update and weight update.

1) *Truth update*: Given the i-th worker's weight w_{i-1}^t of last submission and his/her data $\boldsymbol{x}_i^t$, the truth is computed as

$$\boldsymbol{x}_t^* = \frac{\sum_{i=1}^{N_t} w_i^{t-1} \boldsymbol{x}_i^t}{\sum_{i=1}^{N_t} w_i^{t-1}}. \tag{1}$$

2) *Distance update*: Given the estimated truth $\boldsymbol{x}_t^*$, worker's data and his/her last updated distance d_i^{t-1}, the distance is computed as

$$d_i^t = \alpha * d_i^{t-1} + (1-\alpha) * d(\boldsymbol{x}_t^*, \boldsymbol{x}_i^t), \tag{2}$$

where α is the decay rate used to determine the impact of the historical data on current source weights estimation. And the distance function $d(\cdot)$ indicates the deviation of two records. For the continuous data, the distance is the squared loss: $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{m=1}^{M}(x_m - y_m)^2$. And for discrete data, the distance function is the Hamming distance: $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{m=1}^{M} \mathbf{1}(x_m, y_m)$, where $\mathbf{1}(x, y) = 1$ if $x \neq y$ and 0 otherwise.

3) *Weight update*: Given each worker's distance, the weight is updated as

$$w_i^t = \log\left(\frac{\sum_{i'=1}^{N_t} d_{i'}^t}{d_i^t}\right). \tag{3}$$

B. Local Differential Privacy (LDP)

Local differential privacy (LDP) provides strong privacy guarantees for each user while collecting and analyzing data with distributed architecture.

Definition 1 ((ϵ, δ)-Local Differential Privacy). *A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-LDP if and only if for any pairs of input values v and v' in the domain of $\mathcal{M}$, and for any possible output $y \in \mathcal{Y}$, it holds*

$$\mathbb{P}[\mathcal{M}(v) = y] \leq \mathrm{e}^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(v') = y] + \delta,$$

where $\mathbb{P}[\cdot]$ denotes probability and ϵ is the privacy budget. A smaller ϵ means stronger privacy protection, and vice versa. (ϵ, δ)-LDP means that a mechanism $\mathcal{M}$ achieves ϵ-LDP with probability at least $1 - \delta$, where δ is typically small.

To achieve (ϵ, δ)-LDP, one can use Gaussian noise.

Definition 2 (Gaussian Mechanism). *Given a function $f \to \mathcal{R}$ over a data set D, if $\sigma = \Delta_2 f \sqrt{2 \ln 2/\delta}/\epsilon$ and $\mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian random variable, mechanism $\mathcal{M}$ provides the (ϵ, δ)-LDP when it follows*

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \sigma^2),$$

where $\Delta_2 f = \max_{D,D'} \|f(D) - f(D')\|_2$ is the ℓ_2-sensitivity of f.

w-event differential privacy model [10], [11] provides provable privacy guarantee for any event sequence occurring at any window of w time stamps. Similar to (ϵ, δ)-LDP, we have:

Definition 3 (w-Event Privacy). *A mechanism $\mathcal{M}$ satisfies w-event (ϵ, δ)-LDP, if for all sets $S \subseteq Range(\mathcal{M})$ and all w-neighboring stream prefixes S_t, S_t' and all t, it holds that*

$$\mathbb{P}[\mathcal{M}(S_t) \in S] \leq \mathrm{e}^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(S_t') \in S] + \delta,$$

where w-neighboring of two stream prefixes S_t, S_t' means for each $S_t[i], S_t'[i]$ such that $i \in [t]$ and $S_t[i] \neq S_t'[i]$, it holds that $S_t[i], S_t'[i]$ are neighboring, and for each $S_t[i_1], S_t[i_2], S_t'[i_1], S_t'[i_2]$ with $i_1 \leq i_2$, and $S_t[i_1] \neq S_t'[i_1]$ and $S_t[i_2] \neq S_t'[i_2]$, it holds that $i_2 - i_1 \leq w$.

In addition, according to the privacy composition theorems widely used in the design of mechanisms based on LDP, the w-event privacy has the following property:

Theorem 1 (Sequential Composition Theorem). *Let the mechanism $\mathcal{M}$ takes stream prefix S_t as input, and it can be decomposed into t mechanisms $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_t$, each $\mathcal{M}_k$ provides (ϵ_k, δ_k)-LDP. Then $\mathcal{M}$ satisfies w-event (ϵ, δ)-LDP if*

$$\forall i \in [t], \sum_{k=i-w+1}^{i} \epsilon_k = \epsilon, \sum_{k=i-w+1}^{i} \delta_k = \delta. \tag{4}$$

This theorem enables a w-event private scheme to view ϵ, δ as the total available privacy budget in any sliding window of size w, and appropriately allocate portions of it across the time stamps.

V. Proposed Scheme

A. Overview

As is mentioned above, our PPTD mechanism is designed for the streaming data, which is continuously generated and changing with time. Therefore, workers sensing data from environment can submit their data at any time without restriction, and server updates the truth at regular intervals for the data submitted during this period with the iCRH framework. Additionally, in order to preserve privacy, each worker perturbs the data with Gaussian noise corresponding to the chosen privacy budget divided from the global privacy budget. And workers' rewards are also related to the chosen privacy budget, since the incentive mechanism involves the average distance of truth discovery worker participated.

To measure the long-term privacy leakage, we make use of the definition of w-event privacy and sequential composition theorem of it, and worker's global privacy budget is used evenly in any sliding window of size w. Besides, the sum of privacy budget used each submission of the worker shall not exceed the global privacy budget before his/her exit, otherwise the worker's privacy cannot be guaranteed.

When the worker wants to exit, the server settles his/her rewards according to the average distance of truth discovery that the worker has participated. Although the server only knows distances between estimated truths and worker's disturbed data, making use of Gaussian mechanism, the aggregation of historical distances can reflect the noise level and the reliability of original sensory data.

In next two subsections, we present details of our proposed PPTD scheme and the incentive mechanism.

B. PPTD over Streaming Data

When entering the system, each worker selects the corresponding group according to the protection level of each submission he/she wants. By sensing data from environment, the members of each group submit data with the same partial privacy budget ϵ_i^t, i.e., the same Gaussian noise level. While workers can submit their data at any time without restriction, the server updates the estimated truth at the same time interval.

For every worker, the global privacy budget ϵ is used evenly in a time window of size w, which is a specific number of submissions rather than objective time for individuals. If the worker has consumed up all the global privacy budget, he/she should exit the system. Otherwise, the worker should bear the risk of privacy disclosure, and the server may obtain sensitive information by statistical analysis of historical data. The detailed procedure of our scheme is shown in **Algorithm** 1.

C. the Incentive Mechanism

At the end of the worker i's participation, the server settles his/her reward R_i according to the following equation:

$$R_i = f(\frac{k}{\sum_{t=1}^{k} \hat{d}_i^{t\ 2}}), \tag{5}$$

where $f(\cdot)$ is a linear monotone increasing function, and the server can design it on demand, while k is the number of historical submissions.

As is mentioned above, for every worker, the sliding window size w is a specific number of submissions rather than objective time. Since our proposed framework is not a synchronous system, the number of submissions varies from worker to worker in the same time interval. Therefore, Eq. 5 also means $R_i = f(k/\sum_{t\in w} \hat{d}_i^{t\ 2})$.

Although it seems that in our incentive mechanism the server just calculates on the historical perturbed distances, the reward also involves other factors actually. The sum also depends on partial privacy budget ϵ_i^t, the sliding window size w and the utility of original sensory data, which can measure the contribution of workers comprehensively. And then we explain and proof it in the Section VI.

Algorithm 1: Differential Privacy-Preserving Truth Discovery over Streaming Data

Input : Sensory data streams $\{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$, the global privacy parameters (ϵ, δ), the sliding window size w and the decay rate α.
Output: The estimated truth set $\boldsymbol{x}_t^*$.
Initialize worker weights $w_i^0 = 1$*, and the worker's accumulated distance* $d_i^0 = 0.00001$;
for *each time period* t **do**
  for *worker* i *who has sensed data this period* **do**
    // $\sigma_i^t = \Delta_2 f\sqrt{2\ln 2/\delta_i^t}/\epsilon_i^t$.
    sends $\hat{\boldsymbol{x}}_i^t = \boldsymbol{x}_i^t + \mathcal{N}(0, \sigma_i^{t\ 2})$ to the server;
    if $\sum_{t\in w} \epsilon_i^t > \epsilon$ *or* $\sum_{t\in w} \delta_i^t > \delta$ **then**
      exits the system.
    end
  end
  server receives N_t workers' data;
  for $i \in N_t$ **do**
    updates the estimated truth $\hat{\boldsymbol{x}}_t^*$ according to Eq. 1;
    updates the distances $\hat{d}_i^t$ according to Eq. 2;
    updates the weights $\hat{w}_i^t$ according to Eq. 3;
  end
end

VI. Analyses and Evaluation

A. Security Analysis

Theorem 2. *The proposed differential privacy-preserving truth discovery over streaming data satisfies w-event (ϵ, δ)-LDP.*

proof: We first prove that the perturbation mechanism of any submission t on any worker i satisfies $(\epsilon_i^t, \delta_i^t)$-LDP. In [15], Dwork *et al.* proved that for $c^2 > 2\ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma \geq c\Delta f/\epsilon$ is (ϵ, δ)-differentially private. Therefore, worker i perturbs the data with Gaussian noise $\mathcal{N}(0, \sigma_i^{t\ 2})$ before every submission, which satisfies $(\epsilon_i^t, \delta_i^t)$-LDP.

Then, suppose worker i submits data in the sliding window size w, and the perturbation mechanisms are denoted as $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_t$, since any $\mathcal{M}_t$ satisfies $(\epsilon_i^t, \delta_i^t)$-LDP as is proved above, according to Eq. 4, the whole procedure of proposed differential privacy-preserving truth discovery over streaming data satisfies w-event (ϵ, δ)-LDP.

B. Utility Analysis

Theorem 3. *Suppose that workers submit streaming data set $\{\{\boldsymbol{x}_i^t\}_{i=1}^{N_t}\}_{t\in w}$, and each data $\boldsymbol{x}_i^t$ is M dimensions corresponding to M objects to be sensed. In worker i's group, the expectation of the mean absolute error between the ground truth $\boldsymbol{x}_t^*$ and the estimated truth $\hat{\boldsymbol{x}}_t^*$ by our proposed mechanism satisfies*

$$\mathbb{E}[MAE(\boldsymbol{x}_t^*, \hat{\boldsymbol{x}}_t^*)] \leq 2\mathbb{E}[|\boldsymbol{x}_i^t - \boldsymbol{x}_t^*|] + \frac{\sqrt{2}\sigma_i^t}{\sqrt{\pi}},$$

where $\mathbb{E}[|\boldsymbol{x}_i^t - \boldsymbol{x}_t^*|]$ means the expectation of distance between the ground truth and the original sensory data before perturbation, and σ_i^t is the standard deviation of Gaussian noise, which is determined by partial privacy parameter $(\epsilon_i^t, \delta_i^t)$, i.e., $\sigma_i^t = \Delta_2 f\sqrt{2\ln 2/\delta_i^t}/\epsilon_i^t$.

proof:

$$
\begin{aligned}
&\mathbb{E}[MAE(\boldsymbol{x}_t^*, \hat{\boldsymbol{x}}_t^*)] \\
&= \mathbb{E}[\frac{1}{M}\sum_{j=1}^{M}|x_{\cdot,j}^t - \hat{x}_{\cdot,j}^t|] \\
&= \mathbb{E}[\frac{1}{M}\sum_{j=1}^{M}|\frac{\sum_{i=1}^{N_t} w_i^t x_{i,j}^t}{\sum_{i=1}^{N_t} w_i^t} - \frac{\sum_{i=1}^{N_t} \hat{w}_i^t \hat{x}_{i,j}^t}{\sum_{i=1}^{N_t} \hat{w}_i^t}|] \\
&= \mathbb{E}[\frac{1}{M}\sum_{j=1}^{M}|\frac{\sum_{i'=1}^{N_t}\sum_{i=1}^{N_t} \hat{w}_{i'}^t w_i^t (x_{i,j}^t - \hat{x}_{i',j}^t)}{\sum_{i'=1}^{N_t}\sum_{i=1}^{N_t} \hat{w}_{i'}^t w_i^t}|] \\
&\le \mathbb{E}[\frac{\sum_{i'=1}^{N_t}\sum_{i=1}^{N_t} \hat{w}_{i'}^t w_i^t (\frac{1}{M}\sum_{j=1}^{M}|x_{i,j}^t - \hat{x}_{i',j}^t|)}{\sum_{i'=1}^{N_t}\sum_{i=1}^{N_t} \hat{w}_{i'}^t w_i^t}] \\
&\le \mathbb{E}[\frac{\sum_{i'=1}^{N_t}\sum_{i=1}^{N_t} (\frac{1}{M}\sum_{j=1}^{M}|x_i^t - \hat{x}_{i',j}^t|)}{N_t^2}] \\
&= \mathbb{E}[|\boldsymbol{x}_i^t - \hat{\boldsymbol{x}}_{i'}^t|] \\
&= \mathbb{E}[|\boldsymbol{x}_i^t + \boldsymbol{x}_t^* - \boldsymbol{x}_t^* - \hat{\boldsymbol{x}}_{i'}^t|] \\
&\le 2\mathbb{E}[|\boldsymbol{x}_i^t - \boldsymbol{x}_t^*|] + \mathbb{E}[|\mathcal{N}(0, {\sigma_i^t}^2)|] \\
&= 2\mathbb{E}[|\boldsymbol{x}_i^t - \boldsymbol{x}_t^*|] + \frac{\sqrt{2}\sigma_i^t}{\sqrt{\pi}}.
\end{aligned}
$$

It is worth noting that $\mathbb{E}[|\boldsymbol{x}_i^t - \boldsymbol{x}_t^*|]$ only depends on average performance of workers' devices and the ambient noise, which is not influenced by our scheme. Obviously, the expectation of error at a certain time relates to the partial privacy parameter $(\epsilon_i^t, \delta_i^t)$ of worker i's group.

And we use RMSE to represent the error between the estimated truth and the ground truth. Then Fig. 2 shows the impact of partial privacy budget ϵ_i^t on the accuracy of results, which is consistent with this expectation: the error of truth discovery results reduces with the increase of ϵ_i^t, sharply in the beginning and slowing down as ϵ_i^t grows.

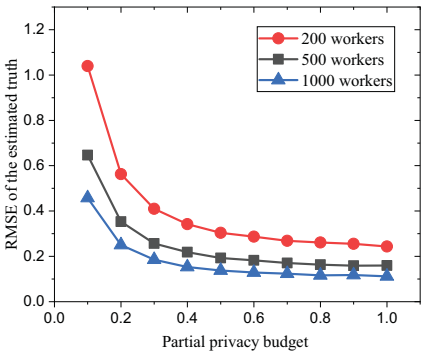

Fig. 2. the Impact of Partial Privacy Budget

C. Effectiveness of Incentive Mechanism

As is mentioned in Section I, due to the limitation of truth discovery algorithm itself, although with the number of participants increasing, the accuracy of the results can be improved (shown in Fig. 2), when this trend reaches a certain extent, the improvement effect is limited, and we use simulation experiments to demonstrate this phenomenon. We set the number of objects to be sensed to 300, the decay rate α to 0.3 and workers from 100 to 900, and the result is shown in Fig. 3. According to the result, a way to further improve accuracy is to reduce the level of noise added to sensory data, which is also the main purpose of incentive mechanism.

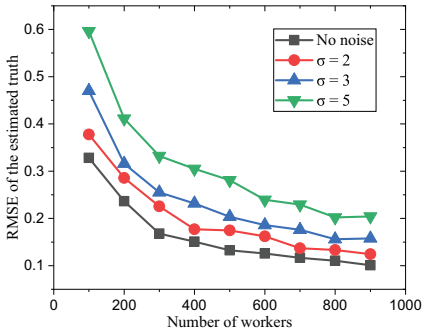

Fig. 3. the Impact of Number of Workers

Theorem 4. *According to the proposed incentive mechanism, for worker, the higher utility the data provided by the worker is, the higher rewards he/she can get, and for server, the more rewards server can provides, the more accurate result of truth discovery can be delivered.*

proof: In Eq. 5, $f(\cdot)$ is a linear monotone increasing function chosen on demand, so we only prove the argument of incentive function relates to partial privacy parameter chosen by worker.

$$
\begin{aligned}
\frac{k}{\sum_{t=1}^{k} {\hat{d}_i^t}^2} &= \frac{k}{\sum_{t=1}^{k} (d_i^t + \mathcal{N}(0, {\sigma_i^t}^2))^2} \\
&= \frac{k}{\sum_{t=1}^{k} {d_i^t}^2 + k\sigma_i^2},
\end{aligned}
\tag{6}
$$

where $\sigma_i = \Delta_2 f\sqrt{2\ln 2/\delta_i^t}/\epsilon_i^t$, and in theory larger privacy budget means higher utility.

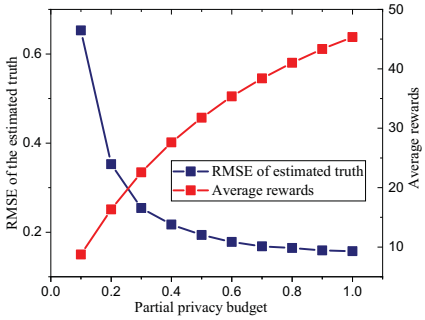

Fig. 4. the Effect of Incentive Mechanism

For server's side, we prove it by simulation experiment. We test it on 500 workers with 300 objects to be sensed, and Fig. 4 looks into the effect of incentive mechanism for both sides of workers and server. It plots the truth error and average rewards of workers with partial privacy budget, which verifies Theorem 4. Additionally, it is worth noting that when partial privacy budget is greater than 0.6, the accuracy has hardly improved but server also pays more for it. Therefore, server should set the groups with partial privacy budget in proper interval otherwise it's not economy for server.

D. Discussion about Incentive Mechanism

We can notice the reward is also influenced by number of submissions k from Eq. 6. According to Eq. 4, on the premise of the global privacy parameter (ϵ, δ) fixed, the larger the sliding window size w means the smaller partial privacy parameter $(\epsilon_i^t, \delta_i^t)$ in every submission i.e., the poorer utility. Therefore, under the condition of the global privacy guaranteed for consecutive periods, workers should try to improve their data utility rather than submit more times in order to make the truth discovery results more accurate.

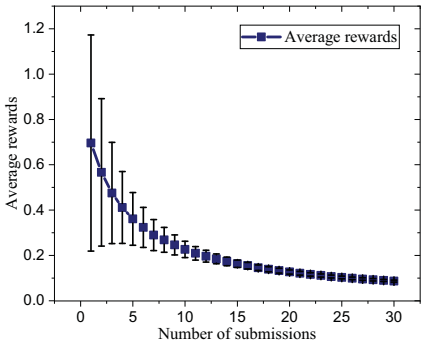

Fig. 5. Rewards with Number of Submissions

Considering the impact of number of submissions k on accuracy, our incentive mechanism in some ways encourages workers use larger partial privacy parameter in each submission though it leads to less submission. Fig. 5 quantifies the relationship between the number of submissions and worker's average rewards each submission. We can see with number of submissions increasing, average rewards reduce. So it's not economy for workers to participate many times. But it is worth noting that the number of submission k is not the less the better, since uncertainty of small amount of submissions is very large, so the worker who spends too large partial privacy budget and less submit data may get a random reward.

VII. Conclusion

In this paper, we proposed an incentive-based differential privacy-preserving truth discovery scheme over streaming data. Considering privacy should not be disclosed in the long-term collection of truth discovery over streaming data, we introduced w-event privacy to measure the privacy protection effect. Besides, on the basis of protecting workers' long-term privacy, we designed a lightweight incentive mechanism to improve the accuracy of truth discovery results. In the future, we will consider how to optimize the use of privacy budget under the condition of limited resources.

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*. ACM, 2014, pp. 1187–1198.

[2] Y. Zheng, H. Duan, and C. Wang, "Learning the truth privately and confidently: Encrypted confidence-aware truth discovery in mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2475–2489, 2018.

[3] C. Miao, L. Su, W. Jiang, Y. Li, and M. Tian, "A lightweight privacy-preserving truth discovery framework for mobile crowd sensing systems," in *Proceedings of the 2017 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2017, pp. 1–9.

[4] 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, 2021.

[5] P. Sun, Z. Wang, Y. Feng, L. Wu, Y. Li, H. Qi, and Z. Wang, "Towards personalized privacy-preserving incentive for truth discovery in crowdsourced binary-choice question answering," in *Proceedings of the 2020 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2020, pp. 1133–1142.

[6] 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, 2022.

[7] Y. Li, C. Miao, L. Su, J. Gao, Q. Li, B. Ding, Z. Qin, and K. Ren, "An efficient two-layer mechanism for privacy-preserving truth discovery," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1705—-1714.

[8] Y. Li, H. Xiao, Z. Qin, C. Miao, L. Su, J. Gao, K. Ren, and B. Ding, "Towards differentially private truth discovery for crowd sensing systems," in *Proceedings of the 40th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2020, pp. 1156–1166.

[9] X. Pang, Z. Wang, D. Liu, J. C. S. Lui, Q. Wang, and J. Ren, "Towards personalized privacy-preserving truth discovery over crowdsourced data streams," *IEEE/ACM Transactions on Networking*, vol. 30, no. 1, pp. 327–340, 2022.

[10] Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 591–606, 2018.

[11] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, "Differentially private event sequences over infinite streams," *Proceedings of the VLDB Endowment*, vol. 7, no. 12, pp. 1155–1166, 2014.

[12] 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.

[13] 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.

[14] K. Xue, B. Zhu, Q. Yang, N. Gai, D. S. L. Wei, and N. Yu, "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.

[15] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.