Towards Personalized Privacy-preserving Truth Discovery over Crowdsourced Data Streams

Journal:	*IEEE/ACM Transactions on Networking*
Manuscript ID	TNET-2020-00376
Manuscript Type:	Original Article
Date Submitted by the Author:	24-Jul-2020
Complete List of Authors:	Pang, Xiaoyi; Wuhan University, School of Cyber Science and Engineering; 1997 Wang, Zhibo; Wuhan University, School of Cyber Science and Engineering Liu, Defang; Wuhan University, School of Cyber Science and Engineering Lui, John C.S; Chinese University of Hong Kong, Computer Science & Engineering; Wang, Qian; Wuhan University, School of Computer Science Ren, Ju; Central South University, School of Information Science and Engineering;
Keywords:	crowdsourcing, truth discovery, privacy preserving, personalization, streaming data

Towards Personalized Privacy-preserving Truth Discovery over Crowdsourced Data Streams

Xiaoyi Pang, Zhibo Wang, *Senior Member, IEEE,* Defang Liu, John C.S. Lui, *Fellow, IEEE,* Qian Wang, *Senior Member, IEEE,* Ju Ren, *Member, IEEE,*

***Abstract*—Truth discovery is an effective paradigm which could reveal the truth from crowdsouced data with conflicts, enabling data-driven decision making systems to make quick and smart decisions, especially for real-time scenarios. However, the increasing privacy concern promotes users to perturb or encrypt their private data before outsourcing, which poses significant challenges for truth discovery in streaming data. Although several privacy-preserving truth discovery mechanisms have been proposed, none of them take personal privacy expectation into consideration. In this work, we propose a novel personalized privacy-preserving truth discovery (PPPTD) framework over crowdsourced data streams to achieve timely and accurate truth discovery while guaranteeing the protection of individual privacy. The key challenges of PPPTD lie in improving the accuracy of truth estimation from the perturbed streaming data with personalized protection level. To address these challenges, we first develop a personalized budget initialization mechanism to quantify each user's privacy protection requirement, and allocate personalized privacy budgets to users according to their privacy requirements. Then we propose a deviation-aware weighted aggregation method to improve the accuracy of truth discovery from streaming data with varying degrees of perturbation. In order to achieve privacy-utility tradeoff, we further propose an influence-aware adaptive budget adjustment mechanism that adaptively re-allocates privacy budgets to users based on the evolution of their influence in the weighted aggregation. We prove that PPPTD can achieve ϵ-differential privacy over the whole data generated by users and satisfy individual personalized privacy requirements. Extensive experiments on two real-world datasets demonstrate the effectiveness of PPPTD.**

***Index Terms*—Crowdsourcing, truth discovery, privacy preserving, personalization, streaming data**

I. Introduction

The ubiquitous mobile devices and the widely used networking technologies have led to the flourish development of crowdsourcing, which can perceive and identify the physical world through the sensing capability of the devices carried by users. The sensory data collected from users can be analyzed to benefit people's daily life in many applications [1]–[3]. A notable issue of crowdsourcing is that the sensory data provided by users are usually noisy, unreliable or inaccurate [4]. Thus, it is challenging to eliminate conflicts among multi-source data and obtain the truthful information.

X. Pang, Z. Wang, D. Liu and Q. Wang are with the School of Cyber Science and Engineering, Wuhan University, 430072, China (e-mail: {xypang, zbwang, defangliu, qianwang}@whu.edu.cn).

J. C.S. Lui is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong (e-mail: cslui@cse.cuhk.edu.hk).

J. Ren is with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China. (e-mail: renju@csu.edu.cn).

Truth discovery attempts to solve the problem with the ability of automatically capturing user quality and accurately inferring reliable information from conflicting data through weighted aggregation. The general principle of truth discovery is that a user is judged to be with high quality if he provides reliable information frequently, and the information is more likely to be true if supported by many users with high quality. Based on this principle, truth discovery can improve the aggregation accuracy [5] from conflicting data, enabling data-driven decision making systems to make smart decisions. Therefore, it has been used in many applications like air quality monitoring [6], social sensing [7], and network quality measurement [8]. Specifically, it is crucial to achieve timely truth discovery for the real-time decision-making systems since only the fresh and truthful data can be helpful. Many truth discovery algorithms [9]–[14] have been proposed ensure the efficiency and accuracy of real-time truth discovery with streaming data. However, due to the threat of individual information disclosure, people become more concerned about their privacy and there is a strong preference that personal data should be protected (e.g., GDPR was launched to regulate the protection of personal data and privacy).

Without a doubt, users have their privacy concerns in truth discovery since the data may contain some sensitive information, which brings forward new demands for the design of truth discovery mechanism with privacy protection guarantee. Some works have achieved truth discovery in the setting of privacy-aware crowdsourcing, which proposed to encrypt or perturb data of each user independently and then aggregate truthful information through these encrypted or perturbed data [15]–[22]. For example, Miao et al. [15] adopted a cloud-based privacy-preserving truth discovery scheme to protect users' sensory data by using threshold Paillier cryptosystem, and performed weighted aggregation on users' encrypted data to obtain truths. Sun et al. [21] proposed a privacy-preserving truth inference method under local differential privacy (LDP), where each user randomizes their answers independently before sending them to the task requester for truth aggregation. Nevertheless, to the best of our knowledge, none of existing works on truth discovery take the different requirements of privacy protection expected by users into account. The fact is that these "one size fits all" approaches are not that applicable to real-world scenarios. For instance, those perturbation-based methods add the same amount of noise to all users' data, which may lead to the situation that some users are overprotected while others are insufficiently protected. Hence, it is necessary to design a new truth discovery mechanism that can satisfy different

privacy expectations of users.

The goal of this paper is to achieve timely and accurate truth aggregation from crowdsouced data streams, and at the same time provide personalized privacy guarantees for individual users. To this end, we have to address two main challenges. The first challenge is: *how to provide personalized privacy protection for each user?* Since it is difficult to measure users' privacy requirements with specific values, it is challenging to quantify users' requirements to decide to what extent they should perturb their data. Another challenge is: *how to accurately find the truth from the perturbed data with different personalized protection levels?* User weights and data submitted by users are two decisive factors of truth weighted aggregation, among which user weights correspond to the quality of users. However, personalized perturbations bring different deviations to user qualities and their submitted data. Thus, it is a big challenge to attenuate the varying effects on user data and quality to achieve timely and accurate truth discovery for crowdsourced data streams.

To address the above challenges, we propose a personalized privacy-preserving truth discovery framework over data streams (PPPTD) to achieve timely and accurate truth discovery while guaranteeing the protection of individual differential privacy. In this framework, users perturb their sensory data with their own personalized privacy budgets according to their privacy protection requirements at each timestamp, and then the perturbed data are collected and our system will infer truthful information through weighted aggregation in time. Specifically, we first propose a *personalized budget initialization mechanism* to quantify each user's privacy requirement and allocate a specific privacy budget that meets the protection requirement to the user. Then we propose a *deviation-aware weighted aggregation mechanism* to accurately infer truths from data with varying degrees of perturbation in time. Moreover, we present an *influence-aware adaptive budget adjustment mechanism* to reallocate privacy budgets to users based on the evolution of their influence in the weighted aggregation process, which allows users with high quality to exert positive influence in the truth computation process so that achieving a trade-off between user privacy and truth accuracy.

Our main contributions are summarized as follows:

- We propose a personalized privacy-preserving truth discovery (PPPTD) framework over data streams. To the best of our knowledge, this is the first work that takes the individual privacy protection requirements into account in the truth discovery process.
- We allocate personalized privacy budgets to users, and develop an influence-aware adaptive budget adjustment mechanism and a deviation-aware weighted aggregation mechanism to achieve accurate inference of truths from the perturbed data submitted by users.
- We prove that PPPTD can provide personalized privacy protection for different users. The extensive experiments on two real-world datasets demonstrate that PPPTD can achieve high accuracy while satisfying ϵ-differential privacy.

The rest of this paper is organized as follows. Section II describes existing truth discovery methods, and Section III presents the system model and the problem formulation. Section IV briefly introduces the preliminary knowledge of truth discovery and differential privacy. Section V and Section VI present the proposed PPPTD and its privacy protection analysis, respectively. Section VII shows the performance evaluation and Section VIII concludes the paper.

II. Related Work

Truth discovery has been greatly developed and applied in recent years. In this section, we discuss truth discovery methods for static scenarios, dynamic scenarios, and privacy-aware scenarios, respectively.

At the very begining, the research of truth discovery focused on the field of traditional database. Yin et al. [23] first formally defined the truth discovery problem, and proposed an iterative method-based *TruthFinder* algorithm to find true facts from conflicting information provided by different websites. It determines the true facts by iteratively inferring the probabilities of facts being true and the trustworthiness of websites. In [24], an unsupervised Bayesian probabilistic model for truth finding on numerical data was designed, which can leverage the characteristics of numerical data in a principled manner, and infer the real-valued truth and source quality. [25] gave an optimization-based answer aggregation method for multiple-choice question answering. It estimated participant weights and aggregated answers simultaneously, and used lightweight machine learning techniques to optimize the accuracy of the results. Numbers of works realized that there are various factors that can raise challenges to truth discovery, and tried to improve the accuracy of user quality estimation and truth discovery results under the circumstances. For instance, a new confidence-aware truth estimation scheme was developed in [26], in which the fact that a source might have different degrees of confidence for his/her different observations was considered, and the truth estimation problem was taken as a maximum likelihood estimation problem. Aware that a source may vary in reliability on different topics or domains, [27] and [28] focused on estimating fine grained source reliability and achieving a more precise truth discovery.

As it moves forward, researchers explored truth discovery in some more complex data scenarios such as data streams. An optimization framework was proposed in [29] to infer truths among conflicting sources of heterogeneous data types, and the proposed framework was further adapted for streaming data and large-scale data. In [9], a model named *EvolvT* based on hidden Markov model was proposed for dynamic truth discovery on numerical data, which captured source dependency besides truth transition regularity and source quality, and established an expectation-maximization (EM) algorithm to infer parameters. Yang et al. [10] proposed an iterative-based truth discovery method to dynamically compute source weights over data streams, where the previous source weights could be used to approximately compute the current truths if the truth inference error caused by not changing source weights at certain timestamps was under a threshold. A streaming fact-finder based on expectation maximization (EM) was designed in [11], which can update previous truth

estimates with new arrived data. Zhao et al. [12] took the problem of truth discovery over data streams as a probabilistic inference problem, and proposed algorithms to real-timely infers the truth as well as source quality, which can read the data only one time. Considering quantitative crowdsourcing applications involving big or streaming data, Ouyang et al. [13] proposed parallel and streaming truth discovery algorithms to realize effective and scalable truth discovery through decomposing large-scale truth discovery problem and leveraging online expectation maximization (EM) algorithm. Li at al. [14] developed a novel truth discovery framework for data streams, which incorporated various iterative methods to effectively infer truths, and can adaptively decide the frequency of source weight computation to improve the efficiency.

With respect to privacy concerns of data sources in truth discovery process, Miao et al. [15] put forward a cloud-enabled privacy-preserving truth discovery (PPTD) framework for crowd sensing systems, which protected users' sensory data and reliability scores with homomorphic encryption, and performed weighted aggregation on the encrypted data to accurately inferred truths. Based on [15], lightweight privacy-preserving truth discovery frameworks are studied [16] and [17], where additively homomorphic cryptosystem was adopted to guarantee both strong privacy and reduce the overhead of users. Zhang et al. [30] leveraged homomorphic Paillier encryption to achieve lightweight privacy-preserving truth discovery and applied it in real-life CIoT applications. Liu et al. [31] considered the dropout of workers in mobile crowdsensing system and proposed a real-time privacy-preserving truth discovery framework for crowdsenced data streams based on secure summation aggregation, which can be robust and achieve highly efficient computation and enough accurate truthful information. A balanced truth discovery (BTD) framework was proposed in [32], which satisfied three requirements in IoT: user privacy, data integrity, and limited computational cost by blurring user data and reducing user participation in the truth discovery process. Sun et al. [21] presented privacy-preserving truth inference method with local differential privacy guarantee, where the truths were inferred from the perturbed answers uploaded by workers. In view of the challenge brought by answer sparsity, a new matrix factorization algorithm is designed to achieve the balance between privacy and utility. These methods treated all users equally and provided them with the same level of privacy guarantee. Li et al. [22] proposed a local differential privacy-based efficient privacy-preserving truth discovery method, which allowed users to add personalized noise to their answers, but the amount of noise was determined by the sampling mechanism. In a word, none of existing methods can provide privacy guarantee according to the personalized privacy requirements of users. This paper aim to achieve personalized privacy-preserving truth discovery by quantifying each user's privacy requirement and adding personalized noise to the sensory data accordingly.

III. Problem Definition

In this section, we first describe the system model and the threat model, and then formally present the problem to be solved in this paper.

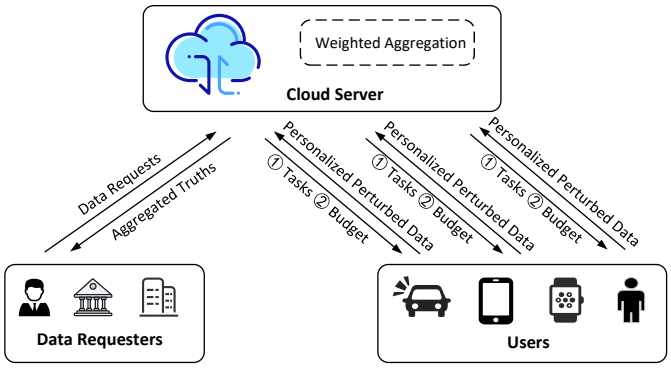


Fig. 1. The system model of PPPTD.

A. Sytem model

The structure of PPPTD is shown in Figure 1, which contains three main entities: data requesters, the server, and users. Data requesters are the customers who send data requests to the server and publish tasks on it. The server is a cloud platform of mobile crowdsourcing (e.g., AMT) that can assign tasks to users, allocate personalized privacy budgets to users according to their privacy protection requirements, and collect users' perturbed sensory data to conduct weighted aggregation to infer the truthful information needed by data requesters. Users are those who carry their mobile devices, and have the ability to perform various sensing tasks assigned by the server and submit the sensory data to the server. It is worth noting that users perturb their sensory data with differential privacy before submitting, and the perturbation levels are controlled by their allocated privacy budgets.

B. Threat Model

We assume that the server and users are curious-but-honest. The server will follow the PPPTD protocol faithfully, but may be curious regarding user individual sensitive information, which means that it may infer some private information of users from the sensory data they submit. Meanwhile, the users will follow the protocol and will not collude with each other, but are likely to deduce the sensory data of others. In this case, users' sensory data should be protected and prevented from being disclosed to any other entity.

C. Problem Formulation

Suppose that data requesters publish N sensing tasks on the server, and there are M users who are interested in these N tasks. At each timestamp $t \in \{1, 2, \cdots, T\}$, these users perform the tasks and provide sensory data for them. Let x_{ij}^t denote the sensory data from the i-th user for the j-th task at timestamp t, then the observation of the i-th user at timestamp t is $X_i^t = \{x_{ij}^t\}_{j=1}^N$, and the observation of all users at at timestamp t is $X^t = \{X_i^t\}_{i=1}^M = \{x_{ij}^t\}_{i=1,j=1}^{M,N}$. The goal of truth discovery is to infer truthful values of all tasks on all timestamps through the weighted aggregation, denoted

by $Z = \{Z^1, Z^2, \cdots, Z^T\}$, where $Z^t = \{Z_j^t\}_{j=1}^N$, and Z_j^t is the truth of task j at timestamp t. However, in order to guarantee personalized privacy protection, each user perturbs his original sensory data according to his privacy protection requirement and only submits the perturbed data to the server. Let $\hat{X}_i^t = \{\hat{x}_{ij}^t\}_{j=1}^N$ denote the submitted perturbed data of the i-th user at timestamp t, and $\hat{X}^t = \{\hat{X}_i^t\}_{i=1}^M = \{\hat{x}_{ij}^t\}_{i=1,j=1}^{M,N}$ is the perturbed data with personalized protection level collected from all users by the server at timestamp t.

In summary, the problem we address in the paper is to accurately infer truthful information Z and estimate user weights W from $\hat{X} = \{\hat{X}^1, \cdots, \hat{X}^T\}$ in time, where $W = \{\{w_i^1\}_{i=1}^M, \cdots, \{w_i^T\}_{i=1}^M\}$, and w_i^t is the weight of user i at timestamp t.

IV. Preliminaries

In this section, we introduce some preliminary knowledge of truth discovery and differential privacy.

A. Truth Discovery

Truth discovery emerges to solve the conflicts among sensory data collected from users, which can automatically estimate source quality from the data in the form of source weights and identify the reliable information (i.e., the truths) among conflicting sources of data. All existing truth discovery mechanisms follow two general principles: a user will be judged to be with high quality if he provides reliable information frequently, and the information will be more likely to be the truth if it is broadly supported by users with high quality. Besides, existing truth discovery mechanisms mainly use a *weighted aggregation* method, which can be summarized as a two-step iterative procedure: *Truth Computation* and *Weight Estimation*. A common process is: truth discovery begins with the initialization of user weights, and then iteratively conducts the truth computation step and weight estimation step until convergence.

Truth Computation: In this step, the user weights are assumed to be known. The truth for the j-th task at timestamp t is calculated based on the following weighted aggregation:

$$Z_j^t = \frac{\sum_{i=1}^M (w_i^t \cdot x_{ij}^t)}{\sum_{i=1}^M w_i^t} \tag{1}$$

Weight Estimation: In this step, the aggregated truths are fixed. The weight of the i-th user at timestamp t is estimated based on the quality of data he provides currently. The closer the data provided by the user is to the aggregated truth, the higher the weight will be assigned to this user. That is:

$$w_i^t = f(\sum_{j=1}^N d(x_{ij}^t, Z_j^t)) \tag{2}$$

where $d(\cdot)$ is a distance function that measures the difference between user-provided data and the aggregated truths, and f is a monotonically decreasing function.

Algorithm 1: Truth Discovery over Crowdsourced Data Streams

Input: Crowdsourced data streams from all users: $\{X^1, X^2, \cdots, X^T\}$
Output: The truths of all tasks at each timesatmp: $\{Z^1, Z^2, \cdots, Z^T\}$

```
Initialize users' weights as W^0 = {w_i^0}_{i=1}^M, for each i ∈ {1,2,···,M}, w_i^0 = 1;
for each timestamp t, t ∈ {1,2,···,T} do
    while the convergence criterion is not satisfied do
        Initialize user weights with W^{t-1};
        for each task j, j ∈ {1,2,···,N} do
            Update the truth z_j^t according to Eq.(1) based on the current estimation of user weights to get Z^t;
        for each user i, i ∈ {1,2,···,M} do
            Update the user weight w_i^t according to Eq.(2) based on the current aggregated truths;
Return {Z^1, Z^2, ···, Z^T} ;
```

In this paper, we adopt the weight estimation of CRH [33] as an instantiation of Eq. (2):

$$w_i^t = -\log(\frac{l_i^t}{\sum_{i'=1}^M l_{i'}^t}) \tag{3}$$

where l_i^t refers to the normalized squared loss function of the i-th user at timestamp t [33], i.e.,

$$l_i^t = \sum_{j=1}^N \frac{(x_{ij}^t - Z_j^t)^2}{std(x_{1j}^t, x_{2j}^t, ..., x_{Mj}^t)} \tag{4}$$

Truth discovery over data streams. Typical truth discovery methods usually conduct iterative procedures of user weight estimation and truth computation on static data. As it moves forward, some truth discovery mechanisms on data streams have been proposed [11]–[14]. In this paper, we consider the real-time crowdsourcing scenarios, so a truth discovery mechanism for crowdsourced data streams is needed. As shown in Algorithm 1, a typical truth discovery mechanism over crowdsourced data streams assumes that the qualities of most users do not change much between two adjacent timestamps, so the user weight at the previous timestamp can be used as the initialized user weight at the current moment. At each timestamp, it begins with the initialization of user weights, and then iteratively conducts the truth computation step and weight estimation step until convergence.

B. ϵ-differential Privacy

Differential privacy tries to prevent individual record in a dataset from being identified.

Definition 1 (ϵ-Differential Privacy [34]): A privacy mechanism $\mathcal{M}$ gives ϵ-differential privacy, where $\epsilon > 0$, if for any

datasets D and $D^{'}$ differing on at most one record, and for all sets $S \subseteq Range(\mathcal{M})$,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D^{'}) \in S] \tag{5}$$

Where the privacy budget ϵ represents the degree of privacy offered by the mechanism, and controls how much noise that should be added to the dataset. In general, a larger perturbation noise is required for a smaller ϵ, which leads to stronger privacy guarantee but worse utility of the dataset.

The Laplace mechanism is the most commonly used mechanism to satisfy ϵ-differential privacy.

Theorem 1 (Laplace Mechanism [35]): Let $f : \mathcal{D} \rightarrow \mathcal{R}^d$, a mechanism $\mathcal{M}$ that adds noise generated independently from a zero-mean Laplace distribution with scale $\triangle(f)/\epsilon$ to each of the output values of $f(D)$ satisfies ϵ-differential privacy, if

$$\mathcal{M}(D) = f(D) + \langle Lap(\triangle(f)/\epsilon)\rangle^d \tag{6}$$

where $\triangle(f)$ is the sensitivity [1] of f.

Now we state two composition properties of differential privacy.

Theorem 2 (Sequential Composition [36]): Let $\mathcal{M}_1$, $\mathcal{M}_2, \cdots, \mathcal{M}_r$ be a set of mechanisms where $\mathcal{M}_i, i \in \{1, 2, \cdots, r\}$ provides ϵ_i-differential privacy. Let $\mathcal{M}$ be another mechanism that executes $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_r$ in sequence and uses independent randomness for each $\mathcal{M}_i$. Then $\mathcal{M}$ satisfies $\sum_i \epsilon_i$-differential privacy.

Theorem 3 (Parallel Composition [36]): Let $Q_1, Q_2, \cdots, Q_\pi$ be the disjoint subsets of dataset Q satisfying $Q = \cup_{i=1}^{\pi} Q_i$ and $Q_i \cap Q_j = \emptyset$ $(\forall\ i \neq j)$. Let $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_\pi$ be a set of mechanisms where $\mathcal{M}_i(Q_i) = f(Q_i) + Lap(\triangle(f)/\epsilon)$ provides ϵ_i-differential privacy. Let $\mathcal{M}(Q) = \cup_{i=1}^{\pi} \mathcal{M}_i(Q_i)$ using independent randomness for each M_i and $f(Q) = \cup_{i=1}^{\pi} f(Q_i)$. Then, $\mathcal{M}(Q)$ satisfies $max\{\epsilon_1, \cdots, \epsilon_\pi\}$-differential privacy.

V. Personalized Privacy-preserving Truth Discovery Mechanism

We propose a personalized privacy-preserving truth discovery mechanism over crowdsourced data, called PPPTD, to real-timely and accurately infer truthful values while providing personalized privacy protection for each user with differential privacy. In this section, we first give a high-level overview of PPPTD, and then introduce the proposed mechanisms in detail.

A. Overview of PPPTD

Figure 2 shows the framework of PPPTD, consisting of the process of perturbation at each user, and the process of personalized budget initialization, influence-aware adaptive budget adjustment and deviation-aware weighted aggregation at the server. The personalized budget initialization mechanism can quantify each user's privacy protection requirement, and allocate a specific privacy budget that meets the requirement of each user. The influence-aware adaptive budget adjustment mechanism can reallocate privacy budget for users based on the evolution of user influence in truth computation. The basic idea of this mechanism is that when the user's influence in the truth computation increases to a certain extent, it is more reasonable to add less perturbation noise to user's sensory data on the premise of satisfying his privacy protection requirement. Only in this way, can the users with high quality greatly exert their positive influence in the truth computation process. The deviation-aware weighted aggregation mechanism can accurately infer truths from the perturbed crowdsourced data with personalized protection levels, in which the impacts of personalized perturbation on weighted aggregation can be eliminated as far as possible. The general process of PPPTD is shown in Algorithm 2. The defined parameters and variables are summarized in Table I.

[1]Please refer to [35] for the definition of sensitivity.

Algorithm 2: PPPTD Algorithm

Input: Crowdsourced data streams from all users: $X = \{X^t\}_{t=1}^{T} = \{X_i^t\}_{i=1,t=1}^{M,T}$

Output: The truths of all tasks at each timesatmp: $Z = \{Z^1, Z^2, \cdots, Z^T\}$

1 The server performs the *personalized budget initialization mechanism* to allocate proper initial privacy budget to each user according to his privacy protection requirement, and get $\epsilon^0 = \{\epsilon_i^0\}_{i=1}^{M}$;

2 **for** *each timestamp t,* $t \in \{1, 2, \cdots, T\}$ **do**

3 The server performs the *influence-aware adaptive budget adjustment mechanism* to reallocate privacy budgets for users whose influence in truth weighted aggregation increase to a certain extent, and get $\epsilon^t = \{\epsilon_i^t\}_{i=1}^{M}$;

4 **for** *each user i,* $i \in \{1, 2, \cdots, M\}$ **do**

5 The user conducts the *personalized perturbation mechansm* to perturb his sensory with the privacy budget assigned to him and submits the perturbed data to the server, $\hat{X}_i^t = \mathcal{M}_i^t(X_i^t) = f(X_i^t) + Lap(\triangle(f)/\epsilon_i^t))$;

6 The server collects $\hat{X}^t = \{\hat{X}_i^t\}_{i=1}^{M}$, and conducts the *deviation-aware weighted aggregation mechanism* on $\hat{X}^t$ until the convergence criterion is satisfied to get Z^t;

7 **Return** $Z = \{Z^1, Z^2, \cdots, Z^T\}$;

B. Personalized Budget Initialization

We aim to provide personalized privacy protection for each user with differential privacy according to their privacy requirements, which means that users should perturb their data at different levels. In differential privacy, a smaller privacy budget means a greater perturbation degree, and will provide stronger privacy guarantee. In turn, when a user has a high privacy protection requirement, a small privacy budget should be allocated to him. Since it is unrealistic for a user to set a specific privacy expectation value in reality, it is difficult to directly map the privacy expectation to a certain privacy budget value. We set up user-friendly instructions for users to enable them to clearly indicate their privacy protection requirements,

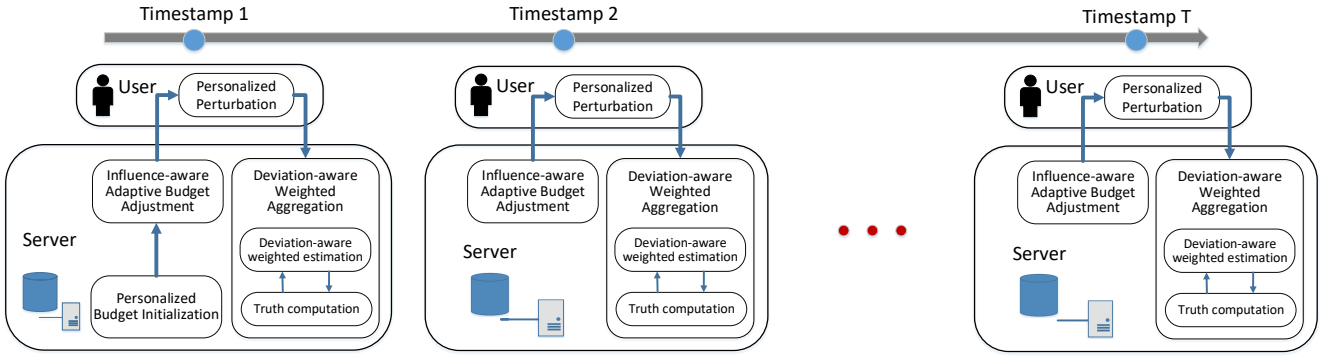


Fig. 2. Framework of the personalized privacy-preserving truth discovery mechanism.

that is, *high*, *medium* and *low privacy requirement*. After the privacy protection requirements of users are clarified, we can quantify them through privacy budgets.

From the practical point of view, we limit the value of user privacy budget to an interval $[\epsilon_0, \epsilon_1]$. Since users' privacy requirements are expressed in three levels (high, medium, and low), it is reasonable to divide the privacy budget into sub-intervals corresponding to these privacy requirement levels. We allocate ϵ_1 to those with low privacy protection requirements as their privacy budgets, and tend to divide the total privacy budget interval into two parts $[\epsilon_0, \epsilon_m)$ and $[\epsilon_m, \epsilon_1)$. Then we let users with high privacy requirements sample their privacy budgets from the first sub-interval, and users with medium privacy requirements sample from the second sub-interval.

The studies in [37], [38] showed that a large majority of users (more than 70% in [37] and 89.3% in [38]) are concerned about privacy leakage arising from the use of their data, and [38] further pointed out that users with high privacy concern are more than those with medium privacy concern, which means a user is more likely to have high privacy protection requirement. In view of the fact that a smaller privacy budget corresponds to a higher privacy guarantee, we assume the privacy budget of the user obeys exponential distribution on the interval $[\epsilon_0, \epsilon_1]$. Thus, the user privacy budget should be sampled from the exponential distribution at the interval that corresponds to his privacy protection requirement.

As we mentioned earilier, the total privacy budget interval can be divided into two sub-intervals $[\epsilon_0, \epsilon_m)$ and $[\epsilon_m, \epsilon_1)$. It is important to find a suitable ϵ_m and an intuitive way is allocating two sub-intervals according to the proportion of users in different privacy budget requirement levels. Suppose that the proportion of users with high and medium privacy protection requirements is α and β respectively, then the fraction of users with low privacy protection requirements is $1-\alpha-\beta$. On an exponential distribution $f(y,\lambda) = \lambda e^{-\lambda y}(y \geq 0)$, we divide the domain of y according to the proportion α, β and then find the mapping relationship between y and ϵ. Suppose we have divided the domain of y into four intervals: $[0, y_0), [y_0, y_m), [y_m, y_1), [y_1, +\infty)$. If we sample a value for the random variable y on the exponential distribution, the probability that it falls on the interval $[y_0, y_m)$, $[y_m, y_1)$, and $[y_1, +\infty)$ should be α, β, and $1-\alpha-\beta-F(y_0)$, respectively, where F is the cumulative distribution function, and $F(y,\lambda) = 1-e^{-\lambda y}(y \geq 0)$. Then we should map y_0 to ϵ_0, map y_m to ϵ_m, and map y_1 to ϵ_1. We get the value of y_m and y_1 by:

$$y_m = Z_{1-\alpha-F(y_0)}, \quad y_1 = Z_{1-\alpha-\beta-F(y_0)} \tag{7}$$

where $Z_{1-\alpha-F(y_0)}$ is a $\{1-\alpha-F(y_0)\}$-quintile that satisfies $P(y > Z_{1-\alpha-F(y_0)}) = 1-\alpha-F(y_0)$, and $Z_{1-\alpha-\beta-F(y_0)}$ is a $\{1-\alpha-\beta-F(y_0)\}$-quintile that satisfies $P(y > Z_{1-\alpha-\beta-F(y_0)}) = 1-\alpha-\beta-F(y_0)$. Given that y_0 is bound to map to ϵ_0, and is generally a very small value that approximately equals to 0, it can be assumed that $F(y_0) = 0$. Therefore, we let

$$y_m = Z_{1-\alpha}, \quad y_1 = Z_{1-\alpha-\beta} \tag{8}$$

After getting the value of y_m, y_1, with the mapping relationship between y and ϵ, we have $y_m/\epsilon_m = y_1/\epsilon_1$. Given a fixed ϵ_1, we can obtain the value of ϵ_m. Then the total privacy budget interval can be divided into two parts: $[\epsilon_0, \epsilon_m), [\epsilon_m, \epsilon_1)$.

Next the server samples a personalized initial privacy budget for each user from the corresponding privacy budget interval. In this way, we can successfully quantify users' privacy requirements through specific privacy budgets, thereby map the privacy protection requirements of users to specific perturbation levels.

C. Influence-aware Adaptive Budget Adjustment

In the scenario of truth discovery over crowdsouced data streams, the quality of a user is not fixed even for the same task, and the user's influence on the final aggregated truth may also change over time. We believe that when the user's influence in truth computation increases to a certain extent, it is reasonable to add less perturbation noise to user's sensory data on the premise of satisfying his privacy protection requirement. For a user with great influence on the computation of truth values, less perturbation on his data leads to less deviation, and enables the user to better exert their positive influence in the truth computation. Therefore, when a user's influence in the truth computation increases over time, it is a logical choice to re-allocate a larger privacy budget to this user while still satisfying his privacy protection requirement. If we do so, users with high influence in the truth computation process would better assist to achieve accurate truth discovery.

We define user i's influence in weighted aggregation at timestamp t as:

$$\zeta_i^t = w_i^t / \sum_{i'=1}^{M} w_{i'}^t$$

TABLE I
NOTATION OF THE VARIABLES

Variable	Description
N	the total number of tasks
M	the total number of users
T	the total number of timestamps
x_{ij}^t	the sensory data from the i-th user for the j-th task at timestamp t
X_i^t	the observation of the i-th user at timestamp t
X^t	the observation of all users at at timestamp t
Z	truthful values of all tasks on all timestamps
Z^t	truthful values of all tasks at timestamp t
Z_j^t	the truth of task j at timestamp t
$\hat{X}_i^t$	the submitted perturbed data of the i-th user at timestamp t
$\hat{X}^t$	the perturbed data with personalized protection level collected from all users by the server at timestamp t
$\hat{x}_{ij}^t$	the submitted perturbed data of the i-th user for the j-th task at timestamp t
W	the set of user weights
w_i^t	the weight of user i at timestamp t
ϵ_0	the lower bound of the privacy budget interval
ϵ_1	the upper bound of the privacy budget interval
ϵ_m	the node that divides the privacy budget interval into two parts
ϵ_i^t	the privacy budget allocated to the i-th user at timestamp t
α	the proportion of users with high privacy protection requirements
β	the proportion of users with medium privacy protection requirements
λ	the parameter of the exponential distribution
ζ_i^t	user i's influence in weighted aggregation at timestamp t
$\triangle w_i^t$	user i's influence evolution at timestamp t
Φ	the unit error of truth aggregation result that caused by the changes of users' influence
$x_{max,j}^v$	the absolute maximum value of j-th task's observations at timestamp t
π	the unit error threshold
Ψ_v^u	the cumulative error of truth aggregation result that caused by the changes of users' influence during timestamp v to timestamp u
$\triangle T$	the maximum period of time where users influence evolution are always less than $\sqrt{\pi}/M$
ρ	the cumulative error threshold
γ	the budget adjustment threshold
$\mathcal{Z}_j$	the truths sequence of the j-th task discovered from the raw data
$\mathcal{Z}_j^t$	the truth of the j-th task discovered from the raw data at timestamp t

Then ζ_i^t may be different for different timestamp t. We aim to capture the evolution of user influence in the truth aggregation, and adaptively adjust the privacy budget allocated to the user according to it. But the challenge is: *when should we reallocate privacy budget to users?* It is unrealistic to update the budget each time when the user influence changes, because it will lead to huge amount of computation, and go against the timely truth discovery. In addition, sometimes the change in user influence may be very small and may not make much difference to the final truth discovery result, so there is no need to update the budget each time when the user influence changes.

Using the similar methodology in [14], we first capture the evolution of user influence, and then measure the changes of truth values caused by the change of the user influence (error) to decide when to reallocate privacy budget for users. If the error is within acceptable limits, namely the changes of users' influence over a period of time has little effect on the truth aggregation result, we can ignore them and do not need to reallocate privacy budgets to users; if the changes of users' influence over a period of time lead to great change in the true aggregation result which exceeds the acceptable limits, then we need to reallocate privacy budgets to these users.

Since user i's influence in the weighted aggregation at timestamp t is $\zeta_i^t = w_i^t / \sum_{i'=1}^M w_{i'}^t$, let $\triangle w_i^t$ denotes user i's influence evolution at timestamp t, which can be computed by:

$$\triangle w_i^t = \zeta_i^t - \zeta_i^{t-1} = w_i^t / \sum_{i'=1}^{M} w_{i'}^t - w_i^{t-1} / \sum_{i'=1}^{M} w_{i'}^{t-1} \tag{9}$$

Let $\Phi = \Phi_{t-1}^t$ $(t \in 1, 2, \cdots, T)$ denote the unit error of truth aggregation result that caused by the changes of users' influence, which is given by:

$$\Phi = \left(\frac{\sum_{i=1}^M |\triangle w_i^t| \cdot x_{ij}^t}{x_{max,j}^t}\right)^2 \tag{10}$$

where $x_{max,j}^t$ is the absolute maximum value of j-th task's observations at timestamp t. Then we have

$$\sqrt{\Phi} = \frac{\sum_{i=1}^M |\triangle w_i^t| \cdot x_{ij}^t}{x_{max,j}^t} = \sum_{i=1}^{M} \frac{|\triangle w_i^t| \cdot x_{ij}^t}{x_{max,j}^t}$$

Since $x_{ij}^t \le x_{max,j}^t$, we have

$$\sqrt{\Phi} \le \sum_{i=1}^{M} |\triangle w_i^t| \tag{11}$$

Given a unit error threshold π, with the Eq. (11), if for each user, the user influence evolution holds $|\triangle w_i^t| \le \sqrt{\pi}/M$, then the unit error $\Phi \le \pi$ is satisfied. That is, the unit error Φ should be no more than π if Eq. (12) is satisfied.

$$|\triangle w_i^t| \le \sqrt{\pi}/M \quad (1 \le i \le M) \tag{12}$$

Let Ψ_v^u denotes the cumulative error of truth aggregation result that caused by the changes of users' influence over a period of time, which is defined as the sum of unit errors in a time period, and it is computed by:

$$\Psi_v^u = \sum_{h=u+1}^{v} \Phi_h^u \tag{13}$$

The maximum value of the cumulative error in a time period under the condition that Eq. (12) holds is:

$$\Psi_v^u \le \triangle T(\triangle T + 1)(2\triangle T + 1)\pi/6 \tag{14}$$

where $\triangle T = v - u$, and the Eq. (14) has been proved in [14].

Assume that timestamp u is the timestamp where user privacy budgets are adjusted, at the beginning it is the timestamp that users are assigned their personalize initial privacy budgets.

The challenge of when to reallocate privacy budgets to users can be tracked by solving the following optimization problem:

$$\begin{aligned} \max \quad & v = u + \triangle T \\ \text{s.t.} \quad & \triangle T(\triangle T + 1)(2\triangle T + 1)\pi/6 \le \rho \\ & |\triangle w_i^h| \le \sqrt{\pi}/M \quad (u \le h \le v, 1 \le i \le M) \end{aligned} \tag{15}$$

where $\triangle T$ is the maximum period of time where users influence evolutions are always less than $\sqrt{\pi}/M$, and the changes of true computation results (the cumulative error) caused by the total changes of users' influence are controlled within a certain range. In other words, during this time period, user influence changes to the maximum acceptable extent. Then it's time to reallocate user privacy budgets.

The user whose total influence evolution $\sum_{h=u}^{v} \triangle w_i^h$ is positive, and influence evolution $\triangle w_i^h$ ($\forall h \in [u, v]$ ($\triangle w_i^h \ge 0 (u \le h \le v)$)) is more than 50% likely to be non-negative is the one that has increasing influence in the truth computation. For these users, we should reallocate privacy budgets at timestamp v. The method is: In the privacy budget interval corresponding to the user's privacy protection requirement, a new privacy budget for the user is obtained by resampling from the interval. Take the i-th user as an example, suppose the adjusted budget allocated to him at timestamp v is ϵ_i^v, then it should satisfy the following constraint:

$$\epsilon_i^v - \epsilon_i^u \le \gamma \tag{16}$$

where ϵ_i^u is user i's last reallocated budget at timestamp u.

In summary, we can determine when to update users' budgets through Eq. (15), and how much to update through Eq. (16).

D. Personalized Perturbation

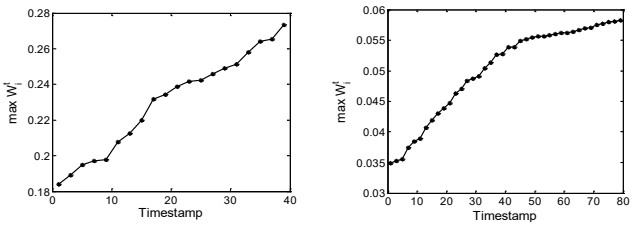
At each timestamp t, each user i perturbs his own sensory data X_i^t with the privacy budget he is allocated to get the perturbed data $\hat{X}_i^t$, calculated by:

$$\hat{X}_i^t = \mathcal{M}_i^t(X_i^t) = f(X_i^t) + Lap(\triangle f/\epsilon_i^t)$$

Since each user has their own personalized privacy budget that corresponds to their privacy protection requirement, users are protected in different levels and the degrees of perturbation on their data are also different. The deviation between each user's original data and submitted data is caused by perturbation, and different degrees of perturbation lead to different levels of deviation. That is, personalized deviation exists in the users' submitted data, which can be formulated as: for $\forall p, q \in \{1, 2, \cdots, M\}$, and $\forall t \in \{1, 2, \cdots, T\}$, if $\epsilon_p^t \ne \epsilon_q^t$, then $|\hat{X}_p^t - X_p^t| \ne |\hat{X}_q^t - X_q^t|$. Then users just submit the perturbed data to the server to from $\hat{X}^t = \{\hat{X}_1^t, \hat{X}_2^t, \cdots, \hat{X}_M^t\}$, which is the total perturbed data with personalized privacy protection level at timestamp t.

E. Deviation-aware Weighted Aggregation

We design a deviation-aware weighted aggregation mechanism to accurately infer truths from the perturbed crowdsourced data with personalized protection levels. At each timestamp t, the server collects all users' submitted data as $\hat{X}^t$, and performs the deviation-aware weighted aggregation on it to get the truths $Z^t = \{Z_1^t, Z_2^t, \cdots, Z_N^t\}$.



(a) *Weather* (max W_i^t) (b) *Intel Lab Data* (max W_i^t)

Fig. 3. The max value of W_i^t

In weighted aggregation method, user weights and user submitted data are two decisive factors of truth, among which user weights correspond to the quality of users. The user's personalized perturbation on his or her data is the key to achieve the personalized privacy protection. However, personalized perturbations bring different deviations to user qualities and their submitted data, leading to greater bias to the aggregated truth values. Only by eliminating these personalized deviations and correcting user weights and user submitted data as much as possible can accurate truth discovery be achieved. Since a large amount of noise leads to the situation that the perturbed data submitted by users diverges greatly from the reality, and causes a huge deviation in user weights. In that case the credibility of users in the weighted aggregation has reduced. Thus, the principle to be followed in the deviation-aware weighted aggregation is that the more noise a user adds to his data, the more his influence in the aggregation progress should be reduced.

Let $W_i^t = \frac{\hat{l}_i^t}{\sum_{i'=1}^{M} \hat{l}_{i'}^t}$, where $\hat{l}_i^t = \sum_{j=1}^{N} \frac{(\hat{x}_{ij}^t - Z_j^t)^2}{std(\hat{x}_{1j}^t, \hat{x}_{2j}^t, \cdots, \hat{x}_{Mj}^t)}$. Based on Eq. (3), the standard way to estimate the user weight through perturbed data should be:

$$w_i^t = -\log W_i^t = -\log \frac{\hat{l}_i^t}{\sum_{i'=1}^{M} \hat{l}_{i'}^t} \tag{17}$$

With the principle mentioned above, we expect a new deviation-aware weight estimation method which can reduce the influence of users who add large amount of noise to their data in the process of weighted aggregation. For that, we find a monotonic decreasing function $g(\epsilon_i^t) = \log_2(3 - \frac{\epsilon_i^t}{\sum_{i'=1}^{M} \epsilon_{i'}^t})$ to revise Eq. (17), and propose that the user weight can be estimated according to:

$$\begin{aligned} w_i^t &= -\log(W_i^t \cdot g(\epsilon_i^t)) \\ &= -\log(\frac{\hat{l}_i^t}{\sum_{i'=1}^{M} \hat{l}_{i'}^t} \cdot \log_2(3 - \frac{\epsilon_i^t}{\sum_{i'=1}^{M} \epsilon_{i'}^t})) \end{aligned} \tag{18}$$

Since $0 < \frac{\epsilon_i^t}{\sum_{i'=1}^{M} \epsilon_{i'}^t} < 1$, we have $1 < \log_2(3 - \frac{\epsilon_i^t}{\sum_{i'=1}^{M} \epsilon_{i'}^t}) < 2$. Empirically we have $0 < W_i^t \le 0.5$, which is demonstrated in Figure 3. Then we have $0 < W_i^t \cdot g(\epsilon_i^t) < 1$, so Eq. (18) is rational.

Suppose that there are two users p and q, whose weights estimated at timestamp t are the same, but user p adds more perturbation noise to his data than user q. That is, $W_p^t = W_q^t$,

and $\epsilon_p^t < \epsilon_q^t$. With this, we have $g(\epsilon_p^t) > g(\epsilon_q^t)$. If we estimate their weights through Eq. (18), we can obtain $w_p^t < w_q^t$. Thus, we can say that Eq. (18) is valid, because with which the influence of the user who adds larger amount of noise to his data is reduced more.

Hence, the truth can be calculated by:

$$Z_j^t = \frac{\sum_{i=1}^{M}(w_i^t \cdot \hat{x}_{ij}^t)}{\sum_{i=1}^{M} w_i^t} \tag{19}$$

Based on Eq.(18) and Eq.(19), we can achieve deviation-aware weighted aggregation, which effectively weakens the personalized deviation caused by personalized perturbation and enables accurate truth discovery.

VI. Theoretical Analysis

In this section, we theoretically analyse the proposed PPPTD mechanism from the perspective of privacy protection.

***Theorem** 4:* PPPTD can provide personalized privacy guarantee for each user.

Proof: Let $\mathcal{M}_i^1, \mathcal{M}_i^2, \cdots, \mathcal{M}_i^T$ be a set of mechanisms where for each $t \in [1, \cdots, T]$, $\mathcal{M}_i^t(X_i^t) = f(X_i^t) + Lap(\triangle(f)/\epsilon_i^t)$ provides ϵ_i^t-differential privacy in isolation. Since $\mathcal{M}_i(X_i) = \{\mathcal{M}_i^1(X_i^1), \mathcal{M}_i^2(X_i^2), \cdots, \mathcal{M}_i^T(X_i^T)\}$, and $\mathcal{M}$ executes $\mathcal{M}_i^1, \cdots, \mathcal{M}_i^T$ in sequence, according to Theorem 2, $\mathcal{M}_i(X_i)$ provides $\sum \epsilon_i^t$-differential privacy.

In PPPTD, for each user i, the total perturbed data he submits to the server satisfies: $\hat{X}_i = \mathcal{M}_i(X_i)$, where $\mathcal{M}_i$ is a perturbation function that satisfies ε_i-differential privacy, and $\varepsilon_i = \sum \epsilon_i^t$. Therefore PPPTD can provide personalized privacy guarantee for each user. ■

***Theorem** 5:* PPPTD satisfies ϵ-differential privacy.

Proof: Let $\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_M$ be a set of mechanisms where $\mathcal{M}_i(X_i) = f(X_i) + Lap(\triangle(f)/\epsilon_i)$ provides ϵ_i-differential privacy. Since $X_1, X_2, \cdots, X_M$ are the disjoint subsets of dataset X satisfying $X = \cup_{i=1}^{M} X_i$ and $X_p \cap X_q = \emptyset$ ($\forall\ p, q \in \{1, 2, \cdots, M\}$ and $p \neq q$), $\mathcal{M}(X) = \{\mathcal{M}_1(X_1), \mathcal{M}_2(X_2), \cdots, \mathcal{M}_M(X_M)\}$. According to Theorem 3, we can get that $\mathcal{M}(X)$ satisfies $max\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_M\}$-differential privacy.

Since the privacy budget assigned to users ranges from ϵ_0 to ϵ_1, we have

$$max\{\varepsilon_1, \cdots, \varepsilon_M\} = max\{\sum_{t=1}^{T} \epsilon_1^t, \cdots, \sum_{t=1}^{T} \epsilon_M^t\} = \epsilon_1 T = \epsilon$$

Thus, $\mathcal{M}$ satisfies $\epsilon_1 T$-differential privacy, namely PPPTD satisfies ϵ-differential privacy, where $\epsilon = \epsilon_1 T$. ■

VII. Performance Evaluation

In this section, we evaluate the performance of PPPTD on real-world datasets to validate its effectiveness.

A. Experiment Settings and Baselines

We conduct experiments on two real-world datasets to compare PPPTD with baseline methods to demonstrate the effectiveness of PPPTD.

Datasets. We use two real-world crowdsourcing datasets: *Weather* dataset [39] and *Intel Lab Data* dataset [40]. *Weather* dataset contains weather data of 30 major USA cities reported by 18 websites every 45 minutes in six days of March 2010. We extract 26 tasks and 16 users from it, and compress the six-day data into 40 timstamps. *Intel Lab Data* contains temperature, humidity, light and voltage data of 54 observation points collected by *Intel Berkeley Research Lab* every 31 seconds from February 28th to April 5th in 2004. We extract 50 tasks and 100 users from it, and we compress the 38-day data into 80 timstamps.

Utility Metric: We use the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) as the utility metric to evaluate the performance of the mechanisms. For any task $j (j \in \{1, 2, \cdots, N\})$, let $Z_j = \{Z_j^1, Z_j^2, \cdots, Z_j^T\}$ denote the sequence of truthful values for task j inferred by the PPPTD mechanism at each timestamp, and $\mathcal{Z}_j = \{\mathcal{Z}_j^1, \mathcal{Z}_j^2, \cdots, \mathcal{Z}_j^T\}$ denote the truths sequence discovered from the raw data. The MAE and MAPE for task j can be computed by:

$$MAE(Z_j, \mathcal{Z}_j) = \frac{1}{T}\sum_{t=1}^{T} |Z_j^t - \mathcal{Z}_j^t| \tag{20}$$

$$MAPE(Z_j, \mathcal{Z}_j) = \frac{1}{T}\sum_{t=1}^{T} |\frac{Z_j^t - \mathcal{Z}_j^t}{Z_j^t}| \tag{21}$$

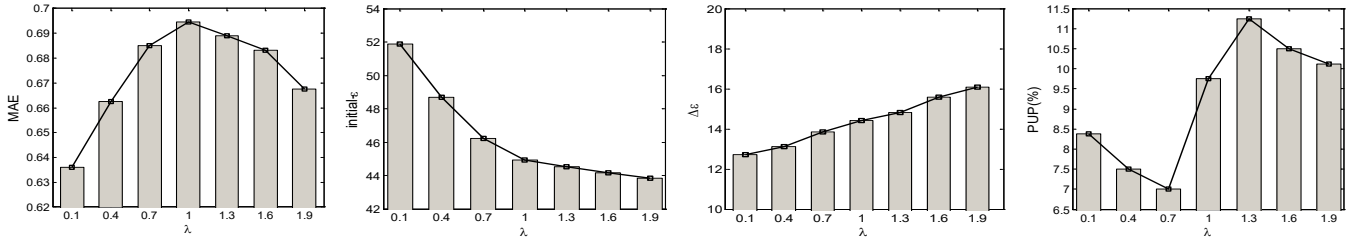
In the experiments, we first calculate the MAE and MAPE for each task and then count up the average of all tasks as the final result of PPPTD.

Personalization Metric: We use the Percentage of Un-Personalized users (PUP) as the personalization metric to evaluate the ability of the mechanisms to provide personalized privacy protection for users. It shows the percentage of users who add the same level of perturbation noise to their data at each timestamp. PUP is realistically the percentage of users whose budgets are updated to the upper bound of the interval with the influence-aware adaptive budget adjustment mechanism. The larger the value of PUP, the less ability of PPPTD to provide personalized privacy protection.

Compared Methods: We first test the effect of each mechanism of the proposed PPPTD. For the influence-aware adaptive budget adjustment mechanism (IAA), we conduct experiments of PPPTD with and without it over two real-world datasets to evaluate the effectiveness of it. For the deviation-aware weighted aggregation mechanism (DWA), we conduct similar experiments to evaluate it's effectiveness.

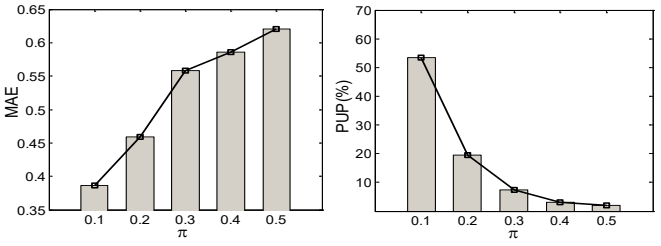
As for the baseline method, we implement a perturbation-based truth discovery method that can meet every user's privacy protection requirement at the same time. For that, all users add the same level of noise to their data with differential privacy before uploading to the server, and the amount of noise must guarantee the highest privacy requirement of users. Then the server conducts weighted aggregation on the perturbed data to infer truths.

Environment: All the mechanisms are implemented in Python, and run on the same machine with 8G RAM, Intel Core i5 processor. We run each experiment 100 times, and report the average results.



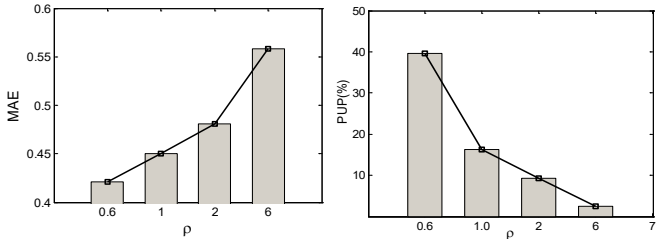
(a) Weather *(MAE)* (b) *Weather* (initial-ϵ) (c) *Weather* ($\triangle\epsilon$) (d) *Weather* (PUP)

Fig. 4. Evaluation on parameter λ.



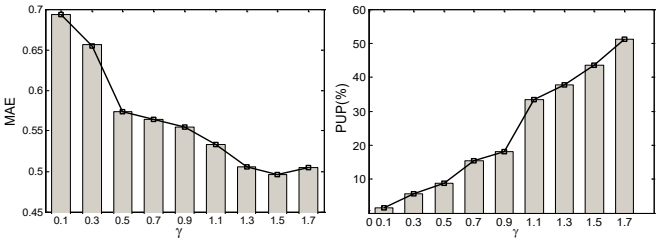
(a) *Weather* (MAE) (b) *Weather* (PUP)

Fig. 5. Evaluation on parameter π.



(a) *Weather* (MAE) (b) *Weather* (PUP)

Fig. 6. Evaluation on parameter ρ.



(a) *Weather* (MAE) (b) *Weather* (PUP)

Fig. 7. Evaluation on parameter γ.

B. Evaluation on Parameters

In this section, we evaluate the effects of parameters λ, π, ρ, and γ on the performance of PPPTD. We test the effect of a parameter over the *Weather* Dataset and *Intel Lab Data Dataset* by changing the value of the parameter while fixing the others. Particularly, we set $\alpha = 0.54, \beta = 0.37$, which are chosen based on findings reported in [38].

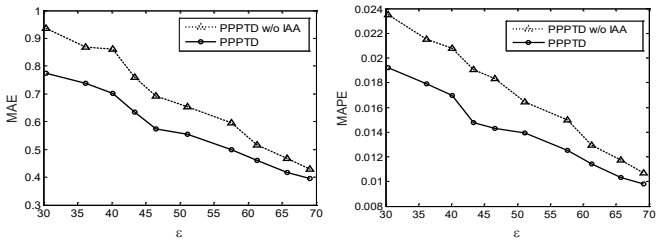
The effect of λ. The value of λ determines the division of privacy intervals, which affects the values of initial budgets allocated to users, as well as the space that user budgets can be updated, thus also the personalization of PPPTD. We evaluate the effect of λ on the utility and personalization of PPPTD, and on the sum of initial budgets and budget update space. For *Weather* Dataset, we fix π to 0.2, ρ to 1, and γ to 1, and make λ varies from 0.1 to 1.9.

Figure 4(a)-(d) shows the evaluation results of the effect of λ on PPPTD over the *Weather* Dataset. As Figure 4(a) shows, MAE first increases and then decreases as λ varies from 0.1 to 1.9. With Figure 4(b) and 4(c), we can explain this state. The value of λ directly affects the value of ϵ_m, that is the division of budget intervals. In PPPTD we first allocate a personalized initial privacy budget to the user from the corresponding budget interval that meets his privacy requirement, and then adaptively adjust the user budget in the same budget interval. So the division of budget intervals can determine the initial budgets allocated to users and the space that user budgets can be updated. Figure 4(b) shows that as λ increases, the sum of initial budgets of users decreases, which means the total noise added to data increases, leading to worse utility. Figure 4(c) shows that the space that user budgets can be updated grows as λ goes from 0.1 to 1.9. Larger update space means that the user budgets have higher probability to be updated to larger values, so users can add less noise to achieve better utility. MAE increases when λ increases from 0.1 to 1 because the sum of initial budgets of users is decreasing and the update space is small. MAE decreases when λ increases from 1 to 1.9 because the update space is large and increasing, and the sum of initial budgets of users only goes down a little bit. Moreover, Figure 4(d) shows influence of different values of λ on PUP, we can learn that PPPTD achieves better personalization when λ is small.

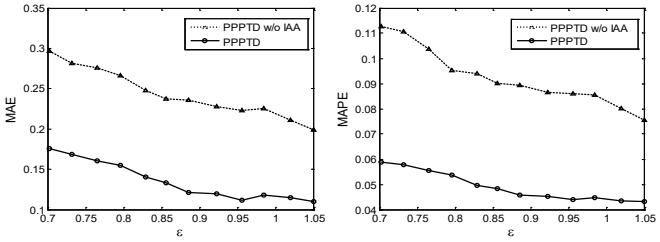
After comprehensive consideration, in the subsequent experiments, we decided to set $\lambda = 0.4$ for the *Weather* Dataset.

The effect of π. The unit error threshold π affects when to update the budgets for which users, and the case of user budgets adjustment is critical to how much noise is added to data totally and how many users can be provided with personalized privacy protection. We evaluate the effect of π on the utility and the personalization of PPPTD. For *Weather* Dataset, we fix ρ to 1000, and γ to 1, and make π varies from 0.1 to 0.5.

The evaluation results of the effect of π over *Weather* Dataset is shown in Figure 5. Figure 5(a) shows the influence of different values of π on MAE. MAE increases as π increases, which is due to the decrease in the frequency of user budget updates. Figure 5(b) shows the influence of different values of π on PUP, and we can find that the smaller the value of π, the larger the PUP. The reason is when the value of π is very small, the user budgets will update very frequently,

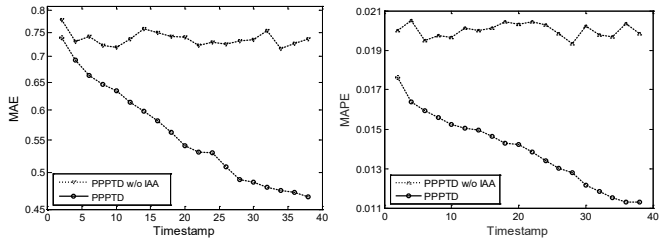


(a) *Weather* (MAE) (b) *Weather* (MAPE)

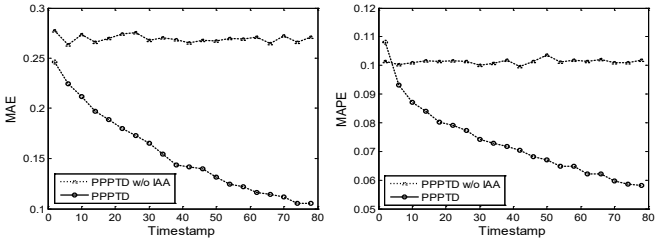


(c) *Intel Lab Data* (MAE) (d) *Intel Lab Data* (MAPE)

Fig. 8. Evaluation on IAA when ϵ changes. ("w/o" stands for "without")



(a) *Weather* (MAE) (b) *Weather* (MAPE)



(c) *Intel Lab Data* (MAE) (d) *Intel Lab Data* (MAPE)

Fig. 9. Evaluation on IAA over time.

thus the budgets of many users are likely to be adjusted to the upper limit of the corresponding interval, leading to poor personalization of PPPTD. In the subsequent experiments, we set $\pi = 0.3$ for the *Weather* Dataset.

The effect of ρ. Similarly, we evaluate the effect of ρ on the noise scale and the personalization of PPPTD. For *Weather* Dataset, we fix γ to 1, and make ρ take values in {0.6, 1, 2, 6}.

The evaluation results of the effect of ρ over *Weather* Dataset is shown in Figure 6, in which MAE increases as ρ increases, and PUP decreases as ρ increases. The reason for this can refer to the above analysis for π since π and ρ have similar effects to PPPTD. In the subsequent experiments, we set $\rho = 1$ for the *Weather* Dataset.

The effect of γ. The adjustment constraint threshold γ controls how much to adjust user budgets. We evaluate the effect of γ on the utility and the personalization of PPPTD.

The results are shown in Figure 7, where MAE decreases and PUP increases when γ varies from 0.1 to 1.7. The reasons are as follows. Smaller γ leads to smaller-scale budget adjustments, which results in only a small reduction in the level of perturbation, and means that users need more updates to reach the upper bound of the interval. Therefore, the smaller the value of γ, the larger the value of MAE, and the smaller the value of PUP. In the subsequent experiments, we set $\gamma = 0.7$ for the *Weather* Dataset.

We also tested the effect of each parameter on the performance of PPPTD over the *Intel Lab Data* Dataset. Due to the space constrain, we do not show the performance results here. As a result, we set $\lambda = 0.4$, $\pi = 0.02$, $\rho = 0.23$, and $\gamma = 0.0008$ for the *Intel Lab Data* Dataset in the subsequent experiments.

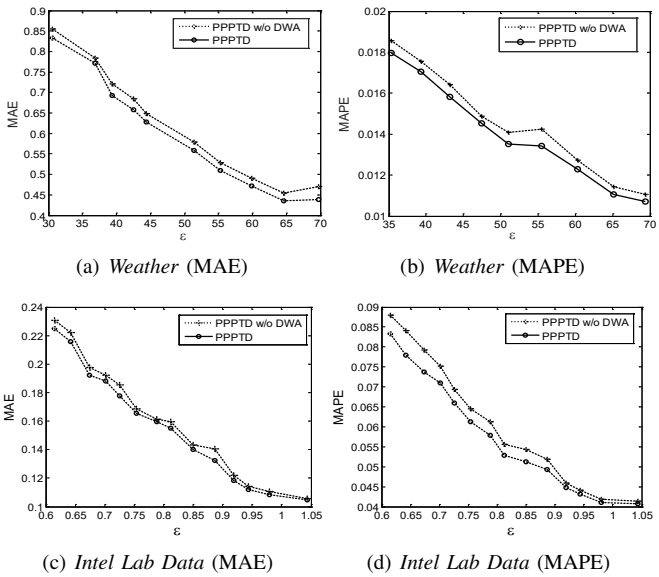
C. Performance Evaluation

In this section, we first test the effect of each mechanism of the proposed method PPPTD, and then compare PPPTD with the baseline method.

The effect of influence-aware adaptive budget adjustment. we conduct experiments of PPPTD with and without the influence-aware adaptive budget adjustment mechanism (IAA) over two real-world datasets to evaluate the effectiveness of the influence-aware adaptive budget adjustment mechanism. Figure 8 shows the result of utility comparison of PPPTD with and withou IAA over two real-world dataset, from which we can find that both MAE and MAPE in PPPTD are smaller than PPPTD without IAA for any given ϵ. Figure 9 further demonstrates that with a fixed ϵ, the MAE and MAPE of PPPTD are always less than those of PPPTD without IAA, and the distance between them becomes larger over time. IAA works more and more over time, thus brings more improvement to utility of PPPTD. It proves that the influence-aware adaptive budget adjustment mechanism (IAA) is useful in improving the utility of PPPTD and it will be more useful as time goes on.

The effect of deviation-aware weighted aggregation. we conduct experiments of PPPTD with and without the deviation-aware weighted aggregation mechanism (DWA) over two real-world datasets to evaluate the effectiveness of the deviation-aware weighted aggregation mechanism. Figure 10 shows the result of utility comparison of PPPTD with and withou DWA. We observe that DWA reduces both MAE and MRE for any given ϵ. Thus, we draw a conclusion that the proposed deviation-aware weighted aggregation mechanism (DWA) does improve the utility of PPPTD.

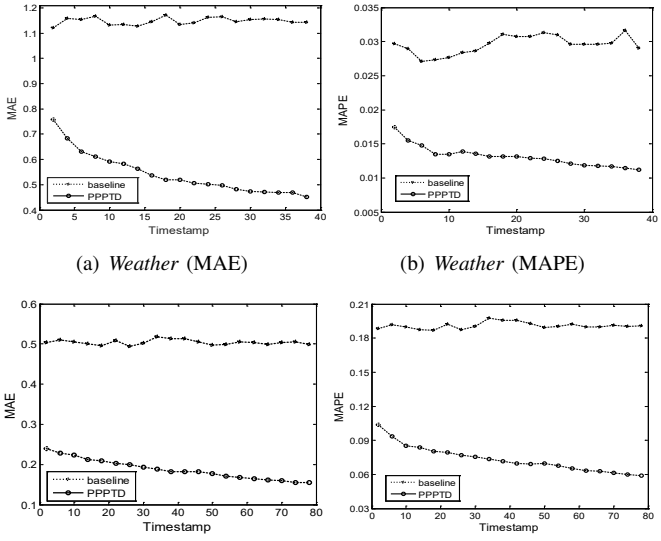
Performance Comparison with the baseline method. Figure 11 shows the results of utility comparison between PPPTD and the baseline method on two real-world datasets. In terms of utility, we can observe that PPPTD overperforms the baseline method for both *Weather* Dataset and *Intel Lab Data* Dataset. This is because PPPTD takes into full consideration the different privacy requirements of users and allows to add different levels of noise to each user's data, so that the total noise is reduced and the data utility is improved. Then we can conclude that PPPTD can provide personalized



(a) *Weather* (MAE) (b) *Weather* (MAPE)

(c) *Intel Lab Data* (MAE) (d) *Intel Lab Data* (MAPE)

Fig. 10. Evaluation on DWA when ϵ changes.



(a) *Weather* (MAE) (b) *Weather* (MAPE)

(c) *Intel Lab Data* (MAE) (d) *Intel Lab Data* (MAPE)

Fig. 11. Performance comparison of PPPTD and the baseline method.

privacy preserving for users and at the same time achieve high accuracy.

VIII. Conclusion

In this paper, we proposed a personalized privacy-preserving truth discovery framework over crowdsourced data streams, called PPPTD, to provide personalized privacy protection for users to meet their personal privacy requirements while real-timely and accurately inferring the truths. In PPPTD, each user is assigned a personalized budget that meets his own privacy requirement, and personally disturbs his data with differential privacy before submitting data to the server. An influence-aware adaptive budget adjustment mechanism and a deviation-aware weighted aggregation mechanism were further proposed for improving the accuracy of inferred truths.

We theoretically proved that PPPTD can provide personalized privacy guarantee for each user meanwhile satisfying differential privacy. The experimental results on two real-world datasets showed that the proposed mechanisms of PPPTD can indeed lead to better utility with a low impact on the overall efficiency, and PPPTD outperforms the baseline method that treats all users equally and does not meet everyone's personalized privacy need.

References

[1] S. Hu, L. Su, H. Liu, H. Wang, and T. F. Abdelzaher, "Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification," *ACM Transactions on Sensor Networks (TOSN)*, vol. 11, no. 4, p. 55, 2015.

[2] S. Liu, Z. Zheng, F. Wu, S. Tang, and G. Chen, "Context-aware data quality estimation in mobile crowdsensing," in *Proc. of IEEE INFOCOM*. IEEE, 2017, pp. 1–9.

[3] Z. Wang, X. Pang, Y. Chen, H. Shao, Q. Wang, L. Wu, H. Chen, and H. Qi, "Privacy-preserving crowd-sourced statistical data publishing with an untrusted server," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1356–1367, 2018.

[4] M. Z. A. Bhuiyan and J. Wu, "Trustworthy and protected data collection for event detection using networked sensing systems," in *Proc. of IEEE SARNOFF*. IEEE, 2016, pp. 148–153.

[5] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *ACM SIGKDD Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

[6] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng, "Truth discovery on crowd sensing of correlated entities," in *Proc. of ACM SenSys*. ACM, 2015, pp. 169–182.

[7] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. of ACM IPSN*. ACM, 2012, pp. 233–244.

[8] Y. Li, J. Gao, P. P. Lee, L. Su, C. He, C. He, F. Yang, and W. Fan, "A weighted crowdsourcing approach for network quality measurement in cellular data networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 300–313, 2016.

[9] S. Zhi, F. Yang, Z. Zhu, Q. Li, Z. Wang, and J. Han, "Dynamic truth discovery on numerical data," in *Proc. of IEEE ICDM*. IEEE, 2018, pp. 817–826.

[10] Y. Yang, Q. Bai, and Q. Liu, "Dynamic source weight computation for truth inference over data streams," in *in Proc. of AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 277–285.

[11] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Proc. of IEEE ICDCS*. IEEE, 2013, pp. 530–539.

[12] Z. Zhao, J. Cheng, and W. Ng, "Truth discovery in data streams: A single-pass probabilistic approach," in *Proc. of ACM CIKM*. ACM, 2014, pp. 1589–1598.

[13] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. J. Norman, "Parallel and streaming truth discovery in large-scale quantitative crowdsourcing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 10, pp. 2984–2997, 2016.

[14] T. Li, Y. Gu, X. Zhou, Q. Ma, and G. Yu, "An effective and efficient truth discovery framework over data streams," in *Proc. of EDBT)*. Springer, 2017, pp. 180–191.

[15] C. Miao, W. Jiang, L. Su, Y. Li, S. Guo, Z. Qin, H. Xiao, J. Gao, and K. Ren, "Cloud-enabled privacy-preserving truth discovery in crowd sensing systems," in *Proc. of ACM SenSys*. ACM, 2015, pp. 183–196.

[16] G. Xu, H. Li, C. Tan, D. Liu, Y. Dai, and K. Yang, "Achieving efficient and privacy-preserving truth discovery in crowd sensing systems," *Computers & Security*, vol. 69, pp. 114–126, 2017.

[17] C. Miao, L. Su, W. Jiang, Y. Li, and M. Tian, "A lightweight privacy-preserving truth discovery framework for mobile crowd sensing systems," in *Proc. of IEEE INFOCOM*. IEEE, 2017, pp. 1–9.

[18] 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," *arXiv preprint arXiv:1810.04760*, 2018.

[19] X. Tang, C. Wang, X. Yuan, and Q. Wang, "Non-interactive privacy-preserving truth discovery in crowd sensing applications," in *Proc. of IEEE INFOCOM*. IEEE, 2018, pp. 1988–1996.

[20] 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.

[21] H. Sun, B. Dong, H. W. Wang, T. Yu, and Z. Qin, "Truth inference on sparse crowdsourcing data with local differential privacy," in *Proc. of IEEE BigData*. IEEE, 2018, pp. 488–497.
[22] 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 *Proc. of ACM KDD*. ACM, 2018, pp. 1705–1714.
[23] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.
[24] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," *Proc. of QDB*, 2012.
[25] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering," in *Twenty-Sixth IAAI Conference*, 2014.
[26] D. Wang and C. Huang, "Confidence-aware truth estimation in social sensing applications," in *Proc. of IEEE SECON*. IEEE, 2015, pp. 336–344.
[27] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proc. of ACM SIGKDD*, 2015, pp. 745–754.
[28] X. Lin and L. Chen, "Domain-aware multi-truth discovery from conflicting sources," *Proc of the VLDB Endowment*, vol. 11, no. 5, pp. 635–647, 2018.
[29] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han, "Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 1986–1999, 2016.
[30] C. Zhang, L. Zhu, C. Xu, K. Sharif, X. Du, and M. Guizani, "Lptd: Achieving lightweight and privacy-preserving truth discovery in ciot," *Future Generation Computer Systems*, vol. 90, pp. 175–184, 2019.
[31] 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.
[32] M. Z. A. Bhuiyan, T. Wang, T. Hayajneh, and G. M. Weiss, "Maintaining the balance between privacy and data integrity in internet of things," in *in Proc. of the 2017 ICMSS*, 2017, pp. 177–182.
[33] 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 *Proc. of ACM SIGMOD*. ACM, 2014, pp. 1187–1198.
[34] C. Dwork, "Differential privacy," in *Proc. of ICALP*, 2006, pp. 1–12.
[35] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006, pp. 265–284.
[36] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proc. of ACM SIGMOD*, 2009, pp. 19–30.
[37] E. Commission *et al.*, "Attitudes on data protection and electronic identity in the european union," *Eurobarometer Special Surveys*, vol. 359, 2011.
[38] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE security & privacy*, vol. 3, no. 1, pp. 26–33, 2005.
[39] L. Berti-Equille and l. Dong, "Weather," 2010. [Online]. Available: http://lunadong.com/fusionDataSets.htm
[40] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel lab data," 2004. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

Xiaoyi Pang received the B.E. degree in Information Security from Wuhan University, China, in 2018. She is currently pursuing her Master degree at School of Cyber Science and Engineering, Wuhan University. Her research interest focuses on privacy protection in mobile crowdsensing system and edge intelligence.

Zhibo Wang received the B.E. degree in Automation from Zhejiang University, China, in 2007, and his Ph.D degree in Electrical Engineering and Computer Science from University of Tennessee, Knoxville, in 2014. He is currently a Professor with the School of Cyber Science and Engineering, Wuhan University, China. His currently research interests include mobile crowdsensing systems, Internet of Things, network security and privacy protection. He is a Senior Member of IEEE and a Member of ACM.

Defang Liu is an undergraduate student of School of Cyber Science and Engineering at Wuhan University, China. He is going to pursue his Master degree at School of Cyber Science and Engineering, Wuhan University. His research interest focuses on privacy protection.

John C.S. Lui received the B.E. degree and Ph.D degree in Computer Science from University of California at Los Angeles, USA. He is currently the Choh-Ming Li Chair Professor in the CSE Department at The Chinese University of Hong Kong. His current research interests are in Internet, network sciences with large data implications, machine learning on large data analytics, network/system security, network economics, large scale distributed systems and performance evaluation theory. John received various departmental teaching awards and the CUHK Vice-Chancellor's Exemplary Teaching Award, and the CUHK Faculty of Engineering Research Excellence Award (2011-2012). He is a Fellow of ACM, IEEE and Hong Kong Academy of Science and Engineering (HKAES).

Qian Wang received the B.S. degree from Wuhan University, China, in 2003, the M.S. degree from Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China, in 2006, and the Ph.D. degree from Illinois Institute of Technology, USA, in 2012, all in Electrical Engineering. He is currently a Professor with the School of Cyber Science and Engineering, Wuhan University. His research interests include wireless network security and privacy, cloud computing security, and applied cryptography. Qian is an expert under "1000 Young Talents Program" of China. He is a co-recipient of the Best Paper Award from IEEE ICNP 2011. He is a Senior Member of IEEE.

Ju Ren received the B.Sc. (2009), M.Sc. (2012), Ph.D. (2016) degrees all in computer science, from Central South University, China. During 2013-2015, he was a visiting Ph.D. student in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Currently, he is a professor with the School of Information Science and Engineering, Central South University, China. His research interests include Internet-of-Things, wireless communication, network computing and cloud computing. He is the recipient of the best paper award of IEEE IoP 2018 and the most popular paper award (2015-2018) of Chinese Journal of Electronics. He currently serves/served as an associate editor for IEEE Transactions on Vehicular Technology and Peer-to-Peer Networking and Applications, and a TPC member of many international conferences including IEEE INFOCOM'19/18, Globecom'17, WCNC'17, WCSP'16, etc. He also served as a poster co-chair of IEEE MASS'18, a track co-chair for IEEE VTC'17 Fall and IEEE I-SPAN'18, and an active reviewer for over 20 international journals.