

## RESEARCH ARTICLE

# Oblivious Statistic Collection With Local Differential Privacy in Mutual Distrust

TAISHO SASADA<sup>1,2</sup>, (Student Member, IEEE), YUZO TAENAKA<sup>1</sup>, (Member, IEEE),  
AND YOUKI KADOBAYASHI<sup>1</sup>, (Member, IEEE)

<sup>1</sup>Graduate School of Science and Technology, Nara Institute Science and Technology, Ikoma 630-0192, Japan

<sup>2</sup>Japan Society for the Promotion of Science, Tokyo 102-0083, Japan

Corresponding author: Taisho Sasada (sasada.taisho.su0@is.naist.jp)


This work was supported in part by the Information-Technology Promotion Agency (IPA)'s Industrial Cyber Security Centre of Excellence (ICS-CoE) Core Human Resources Development Program, and in part by the Japan Society for the Promotion of Science (JSPS)'s KAKENHI Grant JP22J23910.

**ABSTRACT** Location data is valuable for various applications such as epidemiology, natural disasters, and urban planning but causes exposure of sensitive information, e.g., home or work place, from collected data in a datastore. Local Differential Privacy (LDP)-based data collection is a promising technology to protect sensitive information. A mobile device modify data to make each piece of data indistinguishable from others but keep its intrinsic value for statistical characteristics in data. Although LDP fundamentally protects the privacy exposure from a data store, a datastore suffer a shortcomings on it; as a datastore can never validate the modified data due to concealed raw data, that allows anyone to tamper with one's data or inject any amount of data, and thus manipulate the statistics of the whole data in a datastore, called data poisoning attack. As a device does not disclose raw data and a datastore cannot collaborate to validate data with a device who may be an adversary on this mutual distrust relationship, data collection needs an ability to avoid the effect of data poisoning.. The cause of data poisoning is the direct relationship between data volume and statistic; the more data a device sends gives more statistical changes on merged data in a datastore. In this paper, we propose to decouple statistical characteristics from data volumes on LDP-based data collection process to minimize the effect of poisoned data on a datastore. We utilize Oblivious Transfer (OT) protocol to retrieve only statistic characteristics of receiving data at a datastore. As OT protocol inevitably strengthen privacy protection on LDP-based data collection and accordingly drops statistic characteristics of data, We adjust LDP processing to collaboratively work with OT protocol. The proposed adjustment method adapts the protection strength of LDP to OT protocol behavior so that a data store receives data containing sufficient statistical characteristics. We conduct qualitative and experimental overhead analysis and show that our method decouples the relationship between statistical characteristics from data volume. Our experimental result also prove that the overhead can be acceptable on devices such as smartphones and IoT.

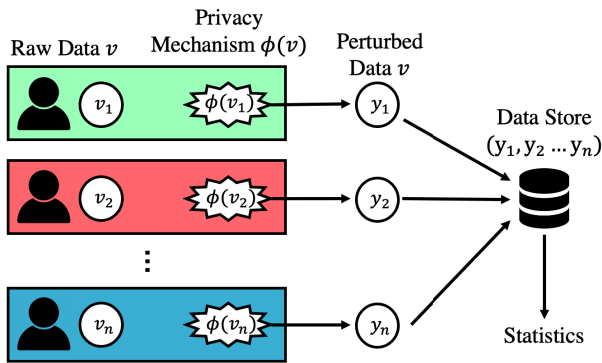
**INDEX TERMS** Local differential privacy, oblivious transfer protocol, location data, privacy-preserving data mining, data security.

## I. INTRODUCTION

The ubiquity of mobile/IoT devices has led the data-driven society. In a data-driven society, we can collect location data through sensors, applications, networks, and APIs. The collected location data can be stored and managed in databases

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Agostino Ardagna .

and the cloud for analysis at any time. One of essential data is location data consisting of people's waypoints and trajectories because it is extremely useful for various applications such as urban planning, epidemiology, and disaster prevention [1], [2], [3], [4]. Although the main interest of those applications is the statistical trend of people's movement, the location data contains sensitive information such as home or work place and could be possibly a cause of privacy exposure.



**FIGURE 1.** The Overview of LDP. Each Device (Client) inputs their raw data  $x$  to a privacy mechanism  $\phi(v)$  that outputs perturbed data  $y$ , and sends  $y$  to the datastore. The datastore computes statistics from the aggregated  $y$ .

To anonymize such sensitive information, Local Differential Privacy (LDP)-based privacy protection attracts great attention [5]. In LDP, a device does not trust any third party, including the datastore or other devices, and will not expose raw data because sharing their raw data can lead to a privacy leak. Instead, the device adds minor changes (or noise) to all values of data at the point of sensing data, i.e. before the device sends their data to a datastore. Figure 1 is the overview of LDP process. Each Device (Client) obtains perturbed data  $y$  via inputting their raw data  $x$  to a privacy mechanism  $\phi(v)$ , and sends perturbed data  $y$  to the datastore. This process is called as perturbation. Since LDP never disclose any raw data outside of the device, no matter how much extra information an adversary may have, he or she cannot be identified. LDP also does not aggregate the raw data in a datastore, client can join the data collection even with untrustworthy datastore.

However, LDP has been pointed out that is vulnerable to various poisoning attacks [6], [7] because LDP has no mechanism for a datastore to validate whether received data is reliable. The cause is that the datastore collects data from devices in the way of correlating data volume and data statistical characteristics. The more data a device sends gives more statistical changes on merged data in a datastore. An adversary's device can tamper with the data or inject arbitrary amounts of data to intentionally change the statistical characteristics of the whole data in the datastore. Since benign devices also do not disclose raw data for fear of privacy leaks, datastore cannot collaboratively validate modified data with a device who may be an adversary. This relationship between devices and datastore create an environment of mutual distrust.

In this paper, we propose a location data collection that extracts the statistical characteristic of receiving data irrespective of its data volume in an environment of mutual distrust. Our method combine LDP and Oblivious Transfer (OT) protocol [8] to obtain statistical characteristics only. OT protocol forces a receiver to discard messages with a certain probability. We utilize this mechanism for letting the datastore (receiver) collect only the statistical characteristic

of data from devices (sender). If a malicious device maliciously crafts the data or amplifies the data volume, OT protocol drop crafted data before reaching datastore. An adversary thus impossibly distort the statistical characteristic and the datastore can mitigate data poisoning attack.

Although OT protocol contributes to decoupling the relationship between statistical characteristics and data volume, we additionally have to adjust the degree of perturbation and whole data volume. The amount of perturbation is calculated on the assumption that all data is received by the datastore in the case of utilizing OT protocol, the datastore receives fewer pieces of data, and thus received data get to have a stronger protection level than expected. Moreover, this sampling by OT protocol makes the amount of data necessary to extract statistics insufficient. Hence, we need to adjust perturbation and data volume for extracting statistical characteristics while keeping all privacy. In our proposal, the devices adjust the degree of perturbation depending on protection strength before LDP processing raw data. After receiving perturbed data, the datastore complements partially missing perturbed data due to OT protocol by generating synthetic data based on a pair of random devices. This generating synthetic data enable to bring the perturbed data volume closer to raw data volume, and extract statistics more accurately.

This work is extension of our preliminary version in Reference [9]. The current version has some significant novelty compared to preliminary version. Our proposal in preliminary version degrade the statistics in data because the method drop almost data by OT protocol; in contrast, our proposal in the current version recovers missing data by oversampling so that statistics can be extracted. The current version enable to extract statistics with accuracy close to that of pure LDP (See Section V). Moreover, we design our current proposal to minimize the impact of two data poisoning attacks. preliminary version did not validate the impact of data poisoning attack, but current version actually show that our method is more secure than pure LDP via extensive experimental evaluation.

The contributions of this study are threefold: First, we establish the location data collection in an environment of mutual distrust by combining LDP and OT protocol. While the mere combination of these two techniques would result in the loss of statistics, we were able to properly transfer the statistics by adjusting the perturbation of LDP and complementing data loss of OT protocol. Second, the proposed method mitigates data poisoning attacks, which have been pointed out as a vulnerability of LDP. Assuming that an adversary is actually included in the data collection in a certain percentage, we conduct an experiment to check what percentage of the statistics can be extracted accurately. The experimental results show that our proposal is more robust against data poisoning attacks than pure LDP. Third, we show that the proposed method can actually collect statistics through experiments on real/synthetic datasets, and measured privacy protection, execution time, and throughput so that the method can be applied to IoT and mobile environments with small memory.

The structure of this paper is as follows: we first give the related work about anonymization, perturbation, and OT-based data collection in Section II. In Section IV, we design a novel LDP-based data collection to decouple statistical characteristics from data volumes. In Section V, we analyze our proposal from a viewpoint of privacy and overhead. In Section VI, we compare our method with related work and discuss the possibilities and limitations of using statistics transfer outside of its scope. Finally, in Section VII, we summarize this study and refer to future work.

## II. RELATED WORK

### A. LDP-BASED LOCATION DATA COLLECTION

LDP enables all device to protect their privacy even device cannot trust the datastore. In LDP, what each device has to do is only adding noise to its own data to ensure indistinguishability and sending numerically different or noisy data. Then, The degree of indistinguishability is determined by the privacy budget  $\epsilon$ . LDP provides mathematical privacy protection regardless of the adversary's background knowledge because each device (client) only sends the perturbed data. Since the datastore cannot check original data in the client device, LDP can guarantee the client's privacy even if the datastore is malicious. To effectively satisfy LDP, Errounda and Liu [10] proposed assigning different privacy protection strengths for each timestamp, and Zhao et al. [11] proposed assigning different privacy protection strengths in time and space within a terminal. In both methods, privacy is strongly protected by LDP. However, LDP is vulnerable to data poisoning attack [6], [7] because LDP has no method for a datastore to verify whether received data is reliable. The adversary thus can succeed in guiding data analysis to the wrong result [6], [7]. To collect only statistical trends from devices (including adversaries), the datastore must receive/discard packets from the devices without violating client's privacy.

### B. OT PROTOCOL-BASED LOCATION DATA COLLECTION

As a way of selecting packets to receive while keeping received data secret, Oblivious Transfer Protocol (OT protocol) [8] is effective technique. In OT protocol, a sender (device) sends many encrypted packets with public key, each of which includes a different piece of data, and a receiver (datastore) decrypts and obtains some of them in a predetermined probability by the trick of key exchange. This protocol is originally used for secure computation, privacy-preserving, etc. OT protocol is a broadly studied cryptographic primitive which involves two mutually distrustful peers who wish to interact with each other in order to transfer messages in an oblivious manner. OT protocol is a two peer protocol between a client and a datastore, by which the client transfers some value to the datastore. Since the OT protocol allows the datastore to unilaterally choose the data it receives, the client has no way of knowing what value the datastore is receiving (decrypting). Many related work have adopted OT protocol is to collect and sample location while protecting

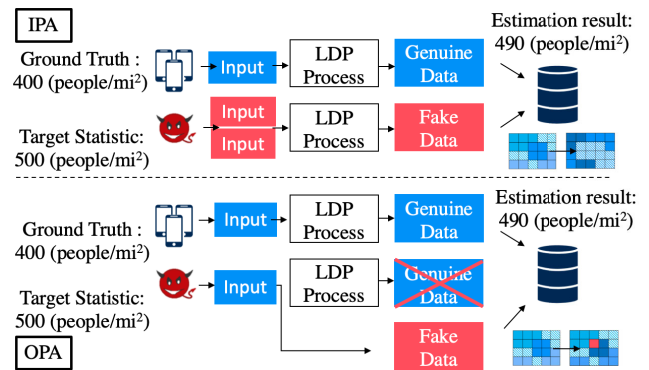


FIGURE 2. Overview of the IPA/OPA.

client's privacy [12], [13], [14]. OT protocol certainly allows sampling of receiving data, but it does not anonymize/perturb the sensitive data itself. For this reason, as with CDP, privacy is not protected if the client does not trust the datastore to protect their privacy while managing the data.

In summary, LDP can protect privacy more securely than other methods (anonymization approaches and CDP), but it is not effective for location data collection if the datastore cannot be trusted by the device, which risks distorting the statistics of the datastore by an adversary. On the other hands, OT protocol allows the datastore to discard/receive data, but the device must unconditionally trust the datastore. To solve this dilemma, we utilize LDP and OT protocol together to collect location data in an environment of mutual distrust.

## III. THREAT MODEL

In this section, we present our threat model. There are two types of attacks on LDP: those at the input stage and those at the output stage [6], [7]. As defined in reference [6], [7], we set the definition of Input Poisoning Attack (IPA) and Output Poisoning Attack (OPA) in location data collection. Figure 2 is the overview of IPA/OPA. In the following subsections, we explain the adversary's capabilities, and motivations in each attacks.

### A. INPUT POISONING ATTACK (IPA)

First of all, we define the adversary's capability in IPA. According to some related work [7], any adversary can easily obtain a large number of fake accounts. We thus assume that the adversary can create fake accounts and manipulate them to amplify the data volume. Specifically, an adversary accesses  $m$  fake accounts and craft their location data and/or sends a large amount of own location information. The datastore extracts the statistical characteristics among the  $n + m$  devices, along with the  $n$  genuine accounts.

The adversary's goal is to increase the amount of data sent using fake accounts, and to distort the statistics that the datastore would have originally obtained from the data. An adversary can distort statistics and disrupt a service that calculates crowding at the landmark (restaurants, railway station, amusement parks, etc) based on location data, thereby degrading the quality of the service. For crowdsourcing

services, distorting statistics cause to spoil the system by crafting fake real-time events (e.g., traffic congestions) in the same way. This type of attack to location crowdsourcing has been pointed out as an example of a practical GPS spoofing in the reference [15].

### B. OUTPUT POISONING ATTACK (OPA)

We assume that an adversary can access a group of fake accounts by illegally registering and/or purchasing accounts from dark markets [7]. If the adversary knows the implementation of the LDP process, he or she can craft the data sent to the datastore by bypassing the perturbation or replacing the process that outputs the perturbed value with a process that outputs an arbitrary value (e.g., using tools to spoof GPS tracking device and/or to amplify data by making  $m$  fake accounts specific locations). OPA is a more serious attack than IPA because the adversary may use it in conjunction with data amplification.

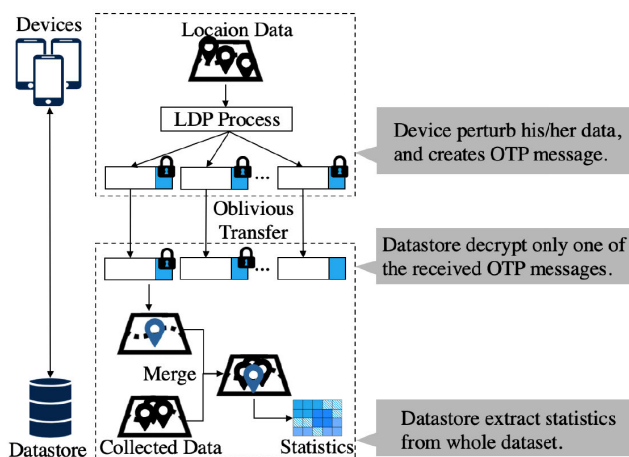
Unlike IPA, we consider that the adversary's goal in OPA is to distort the frequency of certain target locations using fake accounts. To achieve this objective, the adversary carefully spoofs and crafts all the location data of the fake account. As a result, the data collected by the datastore is increased by the number of fake accounts, as well as the number of data, and only certain waypoints or trajectories have a high frequency. By manipulating the frequency of the certain location, the adversary can intentionally manipulate the service. For example, in dating applications, an adversary can measure the distance to a particular user while spoofing his or her own location, and can estimate the approximate location in real-time [16]. Based on this estimation, the adversary can deliberately act in a way that facilitates matching the opponent's device in the system. Not only location data can be leaked, but also cyber-stalking can become real-life, which automated stalking exposing user's privacy even physically. As a real incident of this pattern, there have been incidents where incorrect route recommendations have led tourists to life-threatening deserts with extremely high temperatures and no water supply [15], [17]. In summary, an adversary in OPA uses fake accounts to send spoofed location data to fraudulently increase the frequency of a particular point or route, greatly distorting the statistics that the data store would otherwise obtain. Unlike IPA, OPA is more critical because it allows for statistical poisoning with regard to arbitrary locations.

## IV. OBLIVIOUS STATISTIC COLLECTION

In this section, we describe the proposed Oblivious Statistic Collection. We first describe the overview of the proposed method and limitations in Sect. IV-A. Then, from Sect. IV-B to Sect. IV-D, we describe the solution to each limitations.

### A. LIMITATIONS ON LDP-PROCESSING OVER OT PROTOCOL

To realize the location data collection in an environment of mutual distrust, we combine LDP and OT protocol. The



**FIGURE 3.** The overview of the proposed method. Devices send location data over 1-out-of- $n$  OT protocol, and the datastore merges them to obtain statistical characteristics. Even if there is a malicious device (adversary), they cannot distort the statistics because of the limited data transfer rate.

combination of LDP and OT protocol decouples the relationship between statistical characteristics and data volume, which allowing the collection of statistical features without exposing any raw data outside of the device. Figure 3 is the overview of proposed method, and we explain each process. On the method, the device first samples their data and creates a message to be transmitted using OT protocol. When creating a message, the device adds noise to satisfy the LDP to protect data privacy. All devices samples from the oldest waypoints in order to preserve the continuity of the trajectory. For instance, if the client holds the data volume  $[d]$ , they decomposes  $[d]$  according to the window size  $w_{msg}$  and sends it. The client can send only one waypoint in this transmission method, in other words, our method is 1-out-of- $n$  OT protocol. Even if the adversary amplifies their data volume excessively or spoof their location, their data transfer rate is limited. In short, OT protocol decouples the relationship between statistical characteristic and data volume because the datastore never receives except one data. All devices transfer the created OT protocol messages to the datastore. The datastore receives only one of the messages transferred by OT protocol and merges the received data with the collected data. To prevent the adversary from guiding to the wrong result by data poisoning (amplifying data volume and/or spoofing their data), the datastore receives only one hash value (non-sampled message values are dropped out) over OT protocol. This transfer multiple times enables the datastore to collect only statistical characteristics independent of the device's data volume. Then datastore analyzes the merged data to obtain the statistical characteristics.

However, implementing LDP on OT protocol causes three serious problems: noise amount does not become uniform between each message of OT protocol (Problem 1), noise amount increase due to OT protocol message drops (Problem 2), and statistic loss due to OT protocol message drops

(Problem 3). First, we describe Problem 1. In order to protect privacy in LDP, devices use privacy mechanisms such as the Laplace or exponential mechanism to add noise to individual waypoints. This helps to anonymize the waypoints from each other. The added noise should be consistent to ensure that the overall privacy protection is constant. However, when using the OT protocol, only a portion of the data is transferred. This can result in an inconsistent level of noise in the data received by the datastore, which can lead to either insufficient privacy protection or excessive protection that goes against the device's intent.

Next, we describe Problem 2. Over OT protocol, the sender divides all data into pieces called messages and sends them. The datastore (receiver) receives randomly selected data over OT protocol with a predetermined probability. This mechanism, which is equivalent to sampling, generates excessive noise relative to the amount of data. This is because the device (sender) does not know in advance how much data the datastore will receive, and the device adds noise on the assumption that they will receive all the data. Figure 5 shows a specific example of privacy loss due to pure LDP over OT protocol. In Figure 5, LDP is used on the OT to collect location data while protecting privacy. A simple combination of LDP and OT provides overly strict privacy protection because the majority of data is not received until the datastore samples the message. If data is collected so that  $\epsilon = 10$  for the entire message by protecting it with  $\epsilon = 1$  for each of the 10 messages, but OT protocol actually drops 8 of those messages, the entire message will have  $\epsilon = 2$ , and the spatial correlation would not be maintained correctly. The amount of noise (perturbation) to satisfy the LDP depends greatly on the message length in the OT and how many of the messages are received by the datastore.

Finally, we explain Problem 3. OT protocol divides data into messages and transfers them, but because most of the messages are lost, the data volume received by the datastore is small. For example, if there are 100 participating devices and 100 records of location data are collected per device, pure LDP is able to collect 10,000 records of data, but this method collects only 100 records. Extracting statistics from small-scale data is difficult, and the added noise to satisfy LDP makes it even more difficult to extract statistics when LDP is implemented over OT protocol compared with normal data. Therefore, Combining LDP and OT protocol need data engineering to increase the data volume as the original data so that statistics can be extracted. In this regard, simply increasing the data volume will generate synthetic data that is completely different from the distribution of the original data. In order to extract appropriate statistics, it is necessary to adjust the data volume, taking into account the original data distribution.

## B. ENCODING LOCATION DATA

To solve Problem 1 of Sect. IV-A, we encode location data. By converting from numerical location data (latitude, longitude, timestamp) to categorical location data (hash value),

our method makes the noise amount on each message uniform. By making the amount of noise uniform on a message-by-message basis, the privacy protection strength remains constant even if OT protocol loses messages. Moreover, for a categorical data format (small-domain), the noise amount to satisfy LDP is small [18]. This is because categorical data is more coarse-grained and easier to disambiguate than numerical data, i.e., it is easier to satisfy the LDP.

As encoding method, we use Quadkey [19]. Unlike common encoding for location data such as GeoHash,<sup>1</sup> QuadKey assigns a hash value to each tile based on mercator coordinates rather than latitude and longitude and can represent distances in the real world more accurately. Many research on location data frequently uses quadkey to provide real-world-based services and data analysis [20], [21], [22]. While GeoHash recreates 32 map segments, QuadKey recursively quadrants the map, allowing for the finely controlled collection of location data granularity. In QuadKey, by setting the zoom level  $\Theta_{\text{zoom}}$  to represent the granularity of the segmentation, the datastore can finely control and collect the location data granularity. For each additional level, each tile is divided into four sub-tiles of equal size. In short, QuadKey can more accurately represent geographical distances in the real world. Figure 4 show the example of quadkey's hash around the Trocadéro Square (latitude = 48.858093 and longitude = 2.294694). As illustrated in Figure 4 at the 3-th zoom level ( $\theta_{\text{level}} = 3$ ), Trocadéro Square is mapped into a tile with the quadkey "012". In our proposal, all client (device) use QuadKey and encode their location data into a hash value (categorical data) based on the  $\Theta_{\text{zoom}}$  agreed upon with the datastore.

## C. ADJUSTING PERTURBATION

To address the Problem 2 in Sect. IV-A, our proposal adjusts the total amount of noise added to each message. By adjusting the total amount of noise according to the proportion of messages that are lost, the method prevents excessive privacy protection.

We will describe the order of OT protocol. After encoding to categorical data format by Quadkey, all device perturb their location and create an OT protocol's message for transferring data to the datastore. The devices then add noise to the data (hash value) to protect privacy. As the perturbation for hash value, we use the  $k$ -Ary Randomized Response ( $k$ -RR) [23], which is a perturbation mechanism that outputs a value different from the input value with a certain probability at a discrete value.  $k$ -RR perturbs the data on the device so that it becomes indistinguishable, thus satisfying the  $\epsilon$ -LDP. On  $k$ -RR, the device samples genuine hash value  $v$  on the device with a probability  $p$  of Equation (1) (they sample fake hash value with a probability  $q$  of Equation (1)) to satisfy  $\epsilon$ -LDP,

<sup>1</sup>GeoHash is one of the public domain geocoding methods based on latitude/longitude, developed by Gustavo Niemeyer while creating the geohash.org web service.

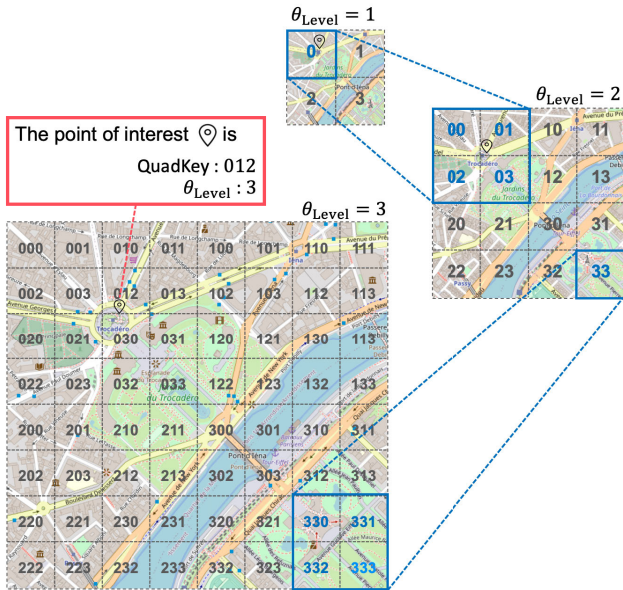


FIGURE 4. QuadKey encoding and zoom level  $\Theta_{\text{zoom}}$ .

and sends part of it to the datastore as a message  $w_{\text{msg}}$ .

$$\mathcal{R}^{\text{kRR}}(y|v) = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + d - 1}, & \text{if } y = v \\ q = \frac{1}{e^\epsilon + d - 1}, & \text{if } y \neq v \end{cases} \quad (1)$$

$$p' = \frac{p_1}{p_2}, q' = \frac{p_2 - p_1}{d - 1} \quad (2)$$

Then, we cause Problem 2 of Sect. IV-A if we just combine k-RR and OT protocol. The proposed method thus adjusts the amount of noise added to location data between the devices and datastore. Specifically, the proposed method perturbs the data of each message by adjusting the protection strength in advance, adapting it to the final protection strength to be achieved (See Figure 6). To determine the number of messages for the entire data, the method calculates the greatest common divisor (GCD). Then, the method employs the Euclidean algorithm, which is the most efficient method to obtain GCD. We derive GCD of  $p$  and message size  $w_{\text{msg}}$  by Euclidean algorithm, and calculate the minimum divisors  $p_1, p_2$  of  $p$  and  $w_{\text{msg}}$  to be k-RR's adjusted probabilities  $p', q'$  after decomposition (See Equation (2)). This allows us to set an appropriate perturbation probability according to  $\epsilon$  and  $w_{\text{msg}}$ . In this case, the perturbation probability  $p'$  after decomposition is less than  $\frac{e^\epsilon}{e^\epsilon + d - 1}$ . Also,  $q'$  is greater than  $\frac{1}{e^\epsilon + d - 1}$ . Therefore,  $p', q'$  are more stringent than  $p, q$  defined in k-RR respectively, and protect privacy more strongly than the pure  $\epsilon$ -LDP.

#### D. AGGREGATION AND OVER-SAMPLING

To handle Problem 3 in Sect. IV-A, our method oversample received data. Compared to pure LDP, the proposed method

receives less data, which making it difficult to correctly obtain statistical characteristics from the aggregated data as a whole.

To increase the amount of data while maintaining statistical characteristics, we increase the sample size by generating synthetic data. We use the Multi-Label Synthetic Minority Over-Sampling Technique (MLSMOTE) [24] for synthetic data generation, which can be used for multi-label categorical data. In order to prevent oversampling from eliminating characteristics in the data (e.g., large cities have higher population densities, rural areas have lower population densities, etc.), we intentionally synthesize the data while preserving imbalance in aggregated data.

Finally, datastore estimates statistical characteristics from the oversampled location data. As an estimator, we use population density data  $(v_i, t_i) (i = 1, 2, \dots, n)$  consisting of device locations to estimate statistical characteristics. This research adopt Kernel Density Estimation (KDE) to estimate the probability density function with relatively high accuracy even when only a small sample are available, and it can match the true probability distribution when the sample is infinite. KDE provides statistical characteristics even for location data which parametric methods cannot be applied due to the difficulty of making distributional assumptions. We define the population density distribution for a given time period as in Equation (3).

$$f(v, t) = \frac{1}{h_t h_S} \sum_{i=1}^I K_t \left[ \frac{t - t_i}{h_t} \right] K_S \left[ \frac{v - v_i}{h_S} \right] \quad (3)$$

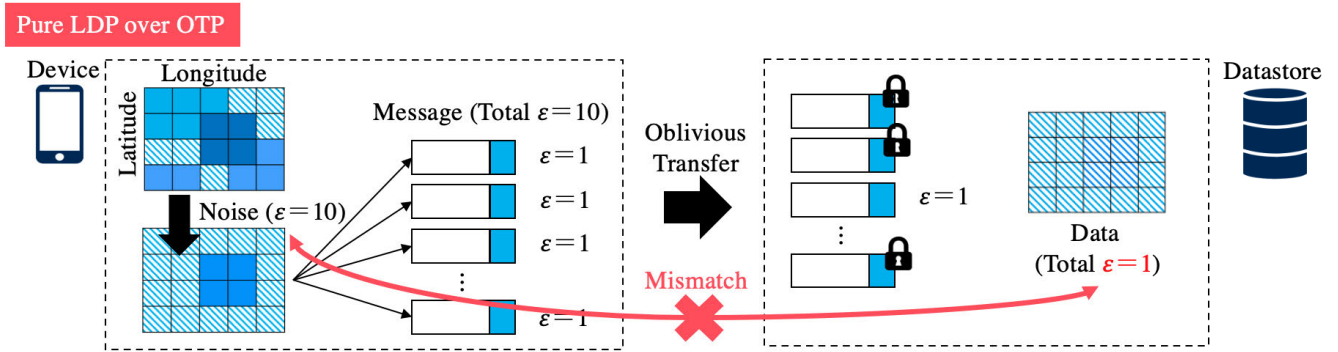
The kernel functions  $K_t$  and  $K_S$  in Equation (3) are the kernel functions for each time and spatial component, where  $K_t$  is the uniform distribution most used in the representation of time and  $K_S$  is the quartic most used in studies of spatial statistics. The bandwidths  $h_t$  and  $h_S$  of the kernel functions  $K_t$  and  $K_S$  are parameters and need to be tuned. In this paper, we tune the parameters by training on population density data and searching for candidate bandwidth pairs. Within the training, the evaluation time point  $t^k (k = 1, 2, \dots, K)$  and the evaluation period  $\Delta t$  is the period of time that is considered to occur at  $t^k$ . For each given pair of candidate bandwidths  $(h_t, h_S)$ , the population density distribution  $c$  at the evaluation point  $t^k$  of a given study period is calculated using the  $m^k$  population density data within the time bandwidth  $d$  in Equation 1, and the set of tiles  $G^{k(b)}$  whose  $f(v, t)$  value is within tile coverage  $\beta$  from the top of all tiles in the area to be forecast is extracted.

## V. EXPERIMENTAL EVALUATION

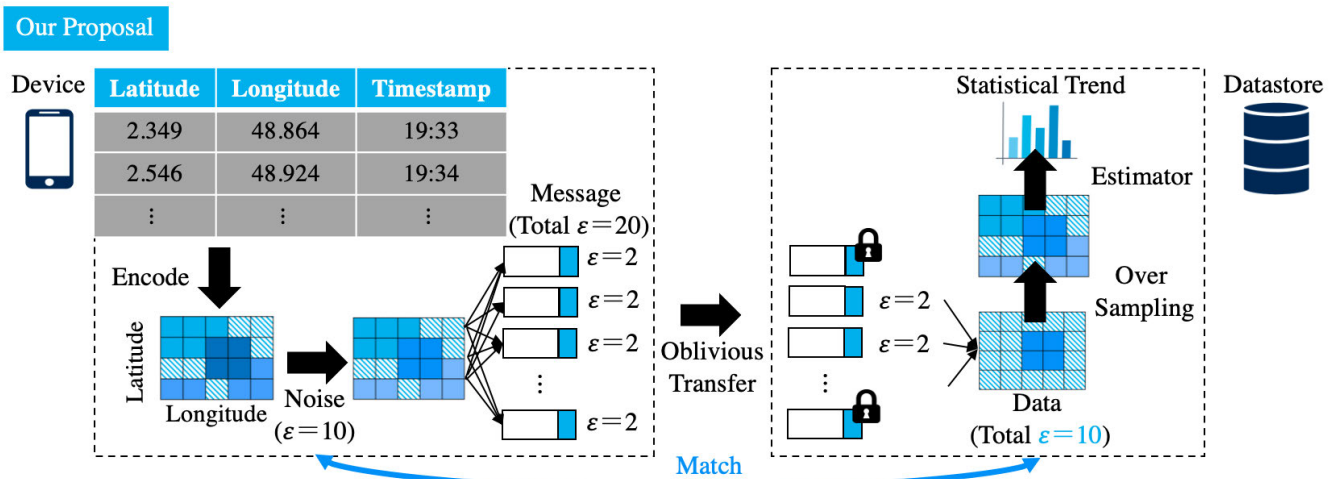
The experimental evaluation measure four aspects: privacy protection, accuracy of statistics collection, robustness to poisoning attacks, and overhead (execution time and throughput).

### A. IMPLEMENTATION

We implement all program on ASRockRack 3U8G+/C621E workstation, CPU is 40-core Intel Xeon Gold 6230 Processor at 2.10 GHz, 262 GB RAM, and the host OS is



**FIGURE 5.** The overview of the combination of pure LDP and OT protocol. Since the datastore (receiver) receives randomly selected data with a predetermined probability over OT protocol, if the device (sender) adds noise to all data, the amount of noise is excessive. Then, the datastore cannot correctly extract statistical characteristics from the data.



**FIGURE 6.** Overview of the perturbation adjustment in our proposed method: to apply LDP on OT protocol, we adjust the Euclidean algorithm to add appropriate noise for the number of messages in OT protocol.

Ubuntu 18.04 LTS. For implementation environment, we use OS-level virtualization Docker to simulate data collection between 100-clients and one datastore. The 100 containers that play the role of client (device containers) were built with the *mem\_limit* provided by docker-compose to limit the size to 2GiB. Device containers send data to datastore containers (datastore containers) using sockets.

As a cryptographic primitives in OT protocol, we adopt Pedersen’s Commitment [25] as a public key scheme, which is a public key exchange scheme that is computationally secure. The client sends the message sequence  $w_{msg}$ , which is perturbed and sampled from the data  $v'$  to the datastore in encrypted form. But the datastore chooses an arbitrary number  $\sigma$  from  $n$ , and decrypt this  $\sigma$ -th value only. Next, the client creates a public key  $K = (g; h)$  using  $w_{msg}$  and  $\sigma$ . From the encrypted values  $w_i$  and  $y_i$ , the datastore calculates  $w_\sigma^b (= v_b)$  using  $(w_i, y_i)$  received from the client and decrypts  $g^{\mu_\sigma} \leftarrow y_\sigma / h^{w_\sigma^b}$ . Since the Decisional Diffie Hellman (DDH) assumption holds for the entire sequence of operations, the

client cannot estimate the  $\sigma$  defined by the datastore, and the datastore likewise cannot know the distribution of the entire message sequence  $w_{msg}$  and the original data  $v$ . Since it has been shown that the combination of Oblivious Transfer and Pedersen’s Commitment is sufficiently secure [26], [27], [28], the client can securely send perturbed data and public keys while protecting their privacy. After decrypting, OT protocol checks the consistency of the last message received (including those that were not decrypted in the datastore). For data authentication, we use Bulletproofs [29] as one of challenge-response authentication to check the consistency. Bulletproofs requires very little communication for verification and can be implemented in low-memory environments such as mobile terminals. Since the client does not have to broadcast the original data over the network, and the content of the challenge sent by the client is different each time, the security risk is very low even if the challenge/response leaks. After data authentication by Bulletproofs, the proposed protocol is terminated.

## B. PRIVACY EVALUATION

The noise amount depend on the privacy budget and perturbation probability. For privacy evaluation, we therefore measure the perturbation probability  $q$  for each privacy budget  $\epsilon$  and compare three of them: pure LDP, LDP over OT protocol, and our proposal. Since the simple combination of OT protocol and LDP results in an excessive amount of noise, we analyze how much probability  $q$  of the all method outputs perturbed value. In this experiment, the client generates location data as random numbers (uniform distribution), serializes them into a sequence, and sends them to the datastore via each method (pure LDP, LDP over OT protocol, and our proposal). Here, we have three parameters: privacy budget and zoom level. The range of the privacy budget is  $\epsilon$  ( $\epsilon \in [0, 10]$ ), and the quadkey's zoom level  $\Theta_{\text{zoom}}$  is verified at  $[1, 4]$  (If the datastore set  $\Theta_{\text{zoom}}$  is 4 between datastore and device, then the tile consist of  $256 (= 4^4)$ ).

The Figure 7 shows the measured  $q$  for each  $\Theta_{\text{zoom}}$  when  $\epsilon$  is varied. A common feature for all  $\Theta_{\text{zoom}}$  is that the difference of  $q$  between pure LDP and our proposal is smaller than the difference between pure LDP and LDP over OT protocol. As explained in Section IV-C, simply combining OT protocol and LDP results in an excessive amount of noise because datastore drop the majority of messages. This is the reason why  $q$  of LDP over OT protocol is abnormally high relative to  $\epsilon$  compared to pure LDP. In Figure 7, LDP over OT protocol has a perturbation probability  $q$  for the privacy budget about 0.1 to 0.2 higher than that of pure LDP. Simply combining OT protocol and LDP did not approximate the perturbation probability of pure LDP. In contrast, the proposed method adjusts the noise amount in advance according to  $\epsilon$  and the number of drop messages. In theory, our perturbation probability  $q'$  approximate  $\frac{1}{e^\epsilon + d - 1}$  (the perturbation probability of pure LDP). Such as LDP over OT protocol, combining multiple privacy protections usually results in too strong privacy protection (removing statistical characteristics), but the proposed method has the same privacy protection probability as pure LDP. In summary,  $q$  of the proposed method is smaller than LDP over OT protocol (the baseline) at all  $\Theta_{\text{zoom}}$ , and the noise amount is as low as pure LDP while achieving data collection in mutual distrust.

Next, we describe the difference among each  $\Theta_{\text{zoom}}$ . When the  $\Theta_{\text{zoom}}$  is low, the coarser granularity of data collection required less noise (small perturbation probability  $q$ ) to guarantee indistinguishability, and there is a large difference among pure LDP, LDP over OT protocol, and our proposal. On the other hands, In the case of  $\Theta_{\text{zoom}}$  is high, the noise amount is large (high  $q$ ) because the finer granularity of the data collection makes it difficult to anonymize data from each other. This is the reason why it is difficult to approximate the same  $q$  of pure LDP. Based on these results, it can be seen that simple LDP over OT protocol without proposed method is not able to collect statistics properly because its privacy protection is too strong.

## C. APPROXIMATION ACCURACY OF STATISTICS COLLECTION

We evaluate whether the location data actually collected by the proposed method loses its statistical characteristics due to privacy protection. Since the proposed method is more private because it combines several privacy protections, it is more difficult to preserve statistics than pure LDP. This experiment evaluates the approximation accuracy of the proposed method against pure LDP.

For the evaluation, we need to set up a specific task using the location data. A popular use of location data in urban development and marketing is the estimation of population density. This study also assumes the use of location data for population density estimation, and evaluates whether the collected data can be used for population density estimation. For measurement, we use Dynamic population distribution dataset for Helsinki Metropolitan Area [30] as real dataset. Not only real dataset, but also we use Power-law/Uniform dataset as synthetic dataset. Since the real dataset also contains correlations and geographic features, we also validate our proposal on a synthetic-dataset which has no correlations and geographic features.

Table 1 indicate the result of approximation accuracy on each dataset. The closer to the accuracy of pure LDP, the better the proposed method is at collecting statistics. The tendency as a whole is that the accuracy is almost linearly proportional to  $\epsilon$ . For each dataset, both pure LDP and our proposal are affected by the imbalance of the dataset. HMA Espoo and HMA Vantaa have scattered distributions (high variance), and even when  $\epsilon = 9$ , the accuracy is not high. On the other hands, HMA Helsinki, HMA Kauniainen and Power-law are locally concentrated in some areas, and thus both pure LDP and our proposal are considered to have achieved high accuracy. Although uniform dataset had a large variance, it was easy to make indistinguishable because the  $\Theta_{\text{zoom}}$  was not large, which lead to high accuracy.

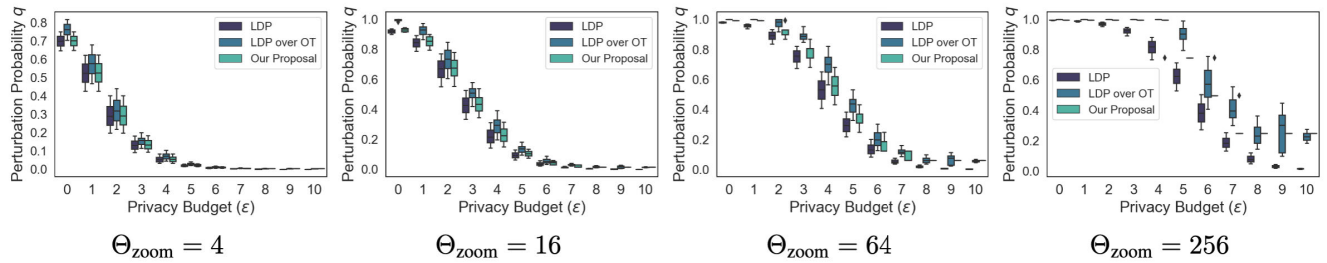
## D. ATTACK EVALUATION

In this section, this study evaluates the robustness of pure LDP and our proposal to poisoning attacks. For our evaluation, we use the uniform dataset in Sect V-B.

### 1) IMPACT OF IPA

Here, we evaluate the robustness pure LDP and our proposal against IPA. By measuring the accuracy of statistic collection when the percentage of adversaries is set to  $[0\%, 25\%, 50\%, 75\%, 95\%]$ , we investigate the degree to which pure LDP and the proposed method suffer deterioration due to IPA. The experimental setup is similar to Sect V-B, with 100 containers of client roles sending data to the datastore. Out of these 100 devices, a certain percentage  $[0\%, 25\%, 50\%, 75\%, 95\%]$  of the adversaries attempt to distort the statistics via the IPA. The adversary sends 10 times more data than the benign device.





**FIGURE 7.** Box plot show the probability  $q$  in privacy mechanism for each privacy budget  $\epsilon$ , measured for each quadkey zoom level  $\theta_{\text{level}}$ . The larger the value of probability  $q$ , the more perturbed the output value is, thus protecting privacy. The smaller the value of  $\epsilon$ , the stronger the privacy protection.

**TABLE 1.** The approximation accuracy of statistical collection.

| Dataset        | Type              | # User | Accuracy (Pure LDP) |              |              |              |              | Accuracy (Our Proposal) |              |              |              |              |
|----------------|-------------------|--------|---------------------|--------------|--------------|--------------|--------------|-------------------------|--------------|--------------|--------------|--------------|
|                |                   |        | $\epsilon=1$        | $\epsilon=3$ | $\epsilon=5$ | $\epsilon=7$ | $\epsilon=9$ | $\epsilon=1$            | $\epsilon=3$ | $\epsilon=5$ | $\epsilon=7$ | $\epsilon=9$ |
| HMA Helsinki   | Real Dataset      | 659k   | 0.182               | 0.523        | 0.739        | 0.810        | 0.929        | 0.180                   | 0.502        | 0.719        | 0.803        | 0.910        |
| HMA Espoo      | Real Dataset      | 297k   | 0.167               | 0.429        | 0.653        | 0.730        | 0.828        | 0.159                   | 0.420        | 0.644        | 0.719        | 0.811        |
| HMA Vantaa     | Real Dataset      | 239k   | 0.149               | 0.332        | 0.648        | 0.781        | 0.839        | 0.148                   | 0.310        | 0.620        | 0.774        | 0.803        |
| HMA Kauniainen | Real Dataset      | 10k    | 0.154               | 0.459        | 0.738        | 0.829        | 0.879        | 0.149                   | 0.445        | 0.711        | 0.801        | 0.825        |
| Power-law      | Synthetic Dataset | 1k     | 0.168               | 0.532        | 0.865        | 0.892        | 0.934        | 0.154                   | 0.510        | 0.836        | 0.878        | 0.912        |
| Uniform        | Synthetic Dataset | 0.1k   | 0.253               | 0.651        | 0.940        | 0.975        | 0.983        | 0.238                   | 0.610        | 0.926        | 0.983        | 0.982        |

Figure 8 (a) to (d) shows the accuracy of population prediction in pure LDP under IPA, and Figure 8 (e) to (h) shows the proposed method. Each percentage represents the ratio of adversaries; the darker the color, the higher the percentage of adversaries. For instance, 25% for the dataset Uniform means that 25 out of 100 users are adversaries. Since pure LDP is not equipped with any mechanism to downsample the received data, the accuracy is strongly affected by IPA, which results in severe degradation. For instance, in Figure 8 (a) to (d), the accuracy is reduced to from 0.2 to 0.4 for all  $\epsilon$ , especially when the ratio of adversaries exceeds 50%, indicating that the statistics are not preserved in the datastore. In contrast, the proposed method limits the data volume per device, no matter how many IPAs an adversary sends. Therefore, our method show that statistics can be preserved as long as the ratio of adversaries is not extremely high. In Figure 8 (e) to (h), even if the ratio of adversaries exceeds 50%, half of the statistics can be preserved compared to the case with no adversaries at all.

2) IMPACT OF OPA

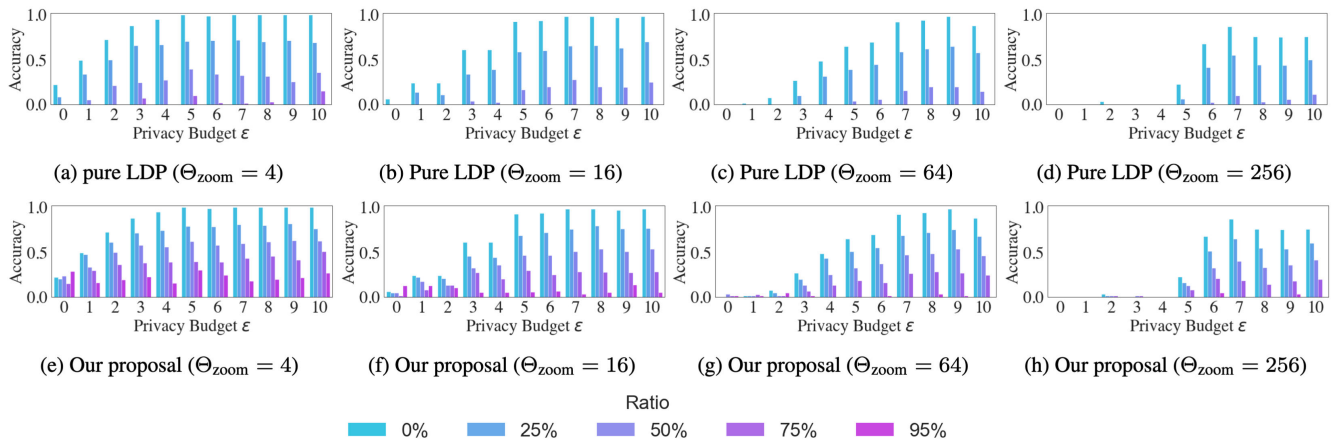
The containers used in the experiment and the ratio of adversaries are exactly the same as in the IPA experiment. Out of these 100 device-role container, a certain percentage [0%, 25%, 50%, 75%, 95%] of the adversaries attempt to distort the statistics via the OPA. Unlike IPA, the adversary intentionally spoof the value of data to manipulate the distribution at an arbitrary location. This experiment define that the adversary sends 10 times as much data as the IPA.

Figure 9 (a) to (d) shows the accuracy of population prediction under OPA in pure LDP, and Figure 9 (e) to (h) shows the proposed method. The higher the value of accuracy

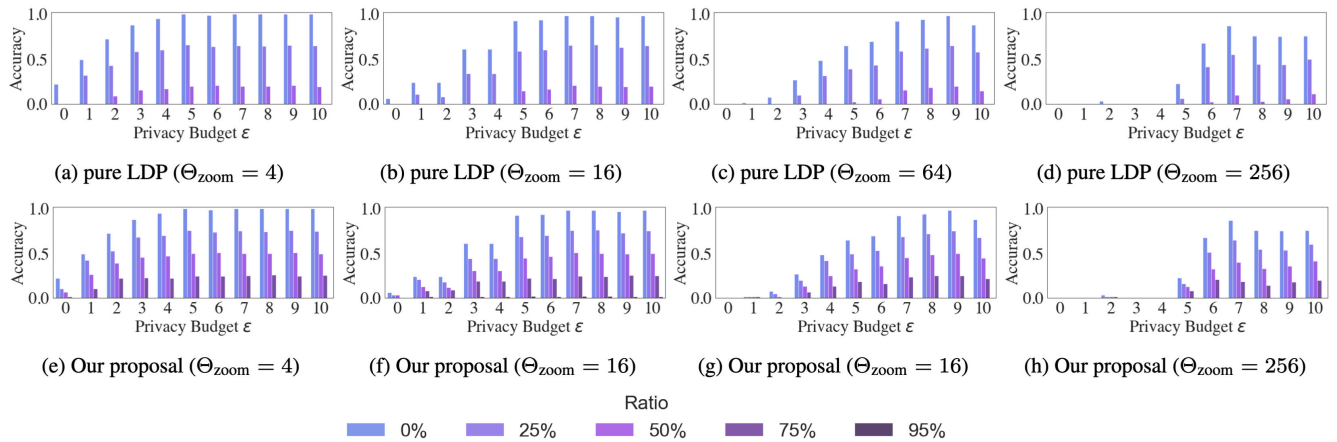
on the vertical axis, the more intrinsic value is preserved on the data store side, and the higher the privacy budget, the more intrinsic value is preserved. Likewise Figure 8 in Sect V-D1, each percentage represents the percentage of adversaries; the darker the color, the higher the percentage of adversaries. Pure LDP is not equipped with any mechanism to limit the data transmission rate, so the accuracy is strongly affected by OPA, which results in severe degradation. In Figure 9 (a) to (d), the accuracy is reduced to from 0.2 to 0.4 for all privacy budgets, especially when the ratio of adversaries exceeds 50%, indicating that the statistics are not preserved in the datastore. In OPA, the adversary intentionally spoofs to a value, which is more severely aggravated than in IPA. In contrast to pure LDP, the proposed method limits the data sent per device. Even if an adversary spoofs and sends values, the impact is only as great as the number of adversarys, since the data volume that can be sent is severely limited. Therefore, our method can preserve statistics as long as the ratio of attackers is not extremely high. In Figure 9 (e) to (h), even if the ratio of adversaries exceeds 50%, our method preserve half of the statistics compared to the case with no adversaries case.

E. OVERHEAD MEASUREMENT

In our method, the execution time and the throughput vary greatly depending on the amount of data and the size of OT protocol message. Depending on these overheads, the proposed method will be difficult to apply in cases where real-time performance is required in data collection and acquisition of statistical characteristics, and in power-saving devices such as IoT and smart devices. To validate their overhead



**FIGURE 8.** The barplot shows the accuracy of statistic collection in pure LDP (upper columns) and proposed method (lower columns) for each privacy budget under IPA.



**FIGURE 9.** The barplot shows the accuracy of statistic collection in pure LDP (upper columns) and proposed method (lower columns) for each privacy budget under OPA.

and discuss the performance, we analyze the overhead by measuring the execution time and throughput.

### 1) EXECUTION TIME

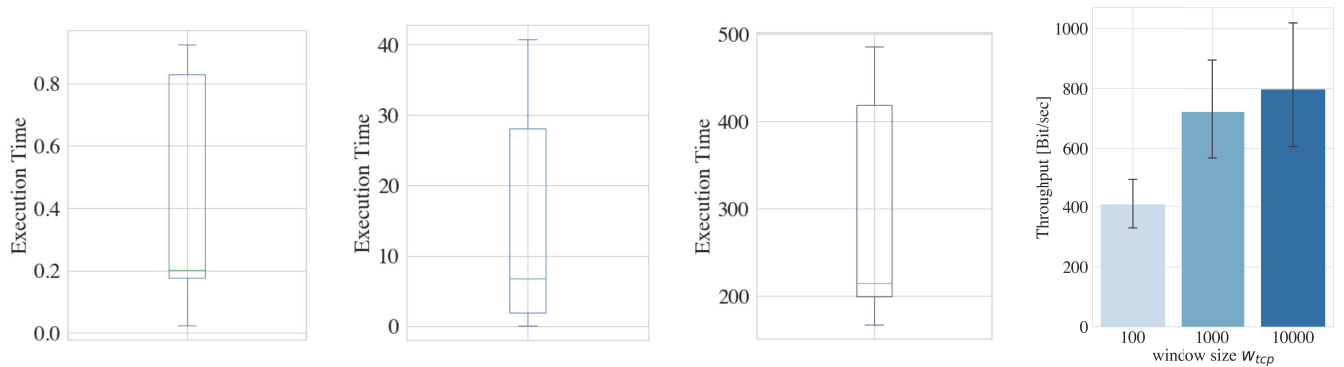
To verify the reality of the proposed method in population density estimation, we evaluate the execution time. Naturally, the execution time varies depending on the size of the message  $w_{msg}$  to be sent. This is because the larger the message size, the more time is required for oversampling and Bullet Proof, which are executed by the datastore after reception. As experimental design, we measure the execution time until all phases of setting parameters ( $w_{msg} : [100, 1000, 10000]$ ), adjusting perturbation for OT protocol, and sending location data between the device and datastore are completed.

All Box plot in Figure 10 shows the execution time distribution of proposed method for each  $w_{msg}$ . The degree of variation is not large for each  $w_{msg}$ , but the execution time itself increases in proportion to the window size. This is because the amount of data over OT protocol increases in proportion to  $w_{msg}$ , and it takes execution time for the datastore

to receive and decrypt the data. Next, we analyze this result in terms of the realistic population density estimation and the requirements. Population estimation is performed at various time slots (e.g., hourly, daily, yearly) [31], [32], [33], [34], but to our knowledge the shortest time slot was every 30 minutes (1800 seconds) [31]. Since even with the largest message size ( $w_{msg} = 10000$ ) in our experiment, the maximum execution time is 486 seconds, the proposed method can be executed in a realistic time for population density estimation.

### 2) THROUGHPUT

In the normal TCP protocol, how much data a datastore can receive at one time depends on the size of data. However, in the proposed method (OT protocol), the data volume received by the datastore is always constant no matter how much data is sent by the device side, thereby the throughput is expected to be constant. As the amount of data processed per unit of time in the datastore, we also measure the receive throughput from the execution time and the amount of data (bits) communicated between the two docker container



**FIGURE 10.** Box plot shows the execution time of the proposed LDP for different window sizes  $w_{msg}$ . The execution time distribution at each  $w_{msg}$  is plotted in combination with the associated box plots. The bar plot shows the throughput at each window size.

(device-role containers and the datastore-role container). Bar plot in Figure 10 shows the throughput for each  $w_{msg}$ . When  $w_{msg}$  is small (such as 100) the variation of the throughput is small, and it is about 150 Bit/sec at most.

However, unlike the intended design of the proposed method, throughput is not constant. The Figure 10 indicate that there is a slight proportional relationship between  $w_{msg}$  and throughput. In particular, there is a significant difference in throughput of up to 650 Bit/sec for  $w_{msg} = 100$  and  $w_{msg} = 10000$ . This is considered to be an indirect effect of the memory load on the datastore container, not the received data itself. The required perturbation probability depends on the privacy budget, zoom level, and window size. The calculation of perturbation probabilities is partially looped by these parameter settings (e.g., the greatest common divisor cannot be obtained by Euclidean Algorithm), resulting in excessive memory load. This is believed to be the reason why the throughput was not constant. This result is reasonable in actual data collection because multiple process threads run in parallel in actual data collection.

## VI. DISCUSSION

Finally, we discuss the comparison with related work, out-of-scope applicability of this research, and our limitations.

### A. OTHER APPLICATION

This study decouple the relationship between statistical characteristics and data volume by combining the LDP over the OT protocol. The Attack on LDP [6], [7] is carried out in various ways, such as generation of malicious raw data, modification of LDP process or parameters, and data amplification. Our proposal may be valid for these attacks. Furthermore, the proposed method provides a guideline that data collection is possible even for mutual-distrust pairs. In our trust model, the device does not trust third parties, including the data store, and thus does not expose any of its original data outside the device. Due to the possibility of LDP attacks, the data store also does not trust that the device will send the correct data at all. In other words, we can say that the proposed method achieves data collection in mutual-distrust pairs. To the best

of our knowledge, there are no studies that have achieved data collection in mutual-distrust pairs. Thereby, applying LDP over the OT protocol may be a solution to these problems in future data collection. Also, the proposed method is not suitable for obtaining precise location information, but it is suitable for collecting landmark-based trajectories (e.g., visiting the Eiffel Tower from Charles de Gaulle airport via the Arc de Triomphe). Landmark-based data collection has been studied mainly for the purpose of congestion, event, location verification, and disaster forecasting [35], [36], [37]. The proposed method that can collect categorical locations is considered to have high affinity.

### B. LIMITATION

The limitations are the integrity of the privacy budget and verification of input data. If the device spoofs the privacy budget after the connection is established, the datastore cannot meet strict LDP. This is very dangerous because it can lead to unintended privacy leaks. Moreover, if the device spoofs the input data to the privacy mechanism, the datastore no longer decouples statistical characteristics from data volume. Spoof detection of input data has long been considered a difficult problem, but it is also necessary in this study.

There may also be cases other than our assumed adversary pattern. In this paper, we assumed that the proportion of adversaries is constant, but in reality it may change as new participants join or leave the data collection. An increase in the proportion of adversaries could also significantly distort the statistics at any given moment. Although we design the proposed method with the 1-out-of-n OT protocol, it may be possible to handle such cases by adjusting the number of messages to be lost, for example, depending on the increase or decrease of the adversary. In that case, however, it would be necessary to design a new function that dynamically adjusts the noise amount as the number of messages lost changes.

## VII. CONCLUSION AND FUTURE WORK

In this study, we designed and implemented LDP over OT protocol to decouple the statistical characteristics from data volume. We proved our proposal is robust to data poisoning

attacks to LDP through experimental evaluation. Our experimental evaluation reveals three facts: (1) The proposed method can extract statistics with higher accuracy than pure LDP, even when strong privacy budgets are set. (2) Pure LDP is vulnerable to both IPA and OPA, but the proposed method is robust and can preserve statistics of location data with high accuracy. (3) The overhead (execution time/throughput) of the proposed method is acceptable for population density estimation from mobile terminals by comparing it with the reference citations.

Next, we summarize our method from the viewpoint of security, sustainability, and efficiency aspects. The proposed method is significantly secure because it uses only sufficiently secure ciphers, and there is no risk of decryption during the OT protocol process. From a security perspective, we mainly discuss whether the proposed method will not leak privacy. Since the proposed method uses only sufficiently secure ciphers, there is no risk of decryption during the OT protocol process. The received data is also authenticated by Bulletproof, so it is secure enough that no intermediary can falsify the data during the step of OT protocol. Next, we discuss sustainability and how far we can respond to changes in devices and datastores. The proposed system can continuously collect data as long as the target device is not damaged. Although there is a limit to the capacity of the data store and data must be discarded when the volume exceeds a certain level, there is no problem from the standpoint of sustainability. Finally, we summarize the efficiency. In the proposed data collection, the datastore is never idle as long as data is sent from the device. The burden on the devices is also small and efficient, consisting only of pre-processing to convert the data into a gridded structure and down-sampling by the OT protocol. However, it is also possible to dynamically control the data volume sent by devices as they move, and proposals for further efficiency improvements are possible.

Finally, we describe some interesting and important directions for future work. It is known that location data does not satisfy strict LDP if data is continuously published. We thereby can consider substituting the privacy mechanism that is assumed to be used for location data. Experiments using actual mobile devices instead of Docker environment would also be of great value and interest. In actual mobile devices, delays occur depending on the throughput of the privacy mechanism, and delays also affect the accuracy of analysis on the datastore. We also suspect there are other applications for this proposal in privacy research beyond population density estimation that could be investigated.

## REFERENCES

- [1] J. Zhang, W. Wang, F. Xia, Y.-R. Lin, and H. Tong, "Data-driven computational social science: A survey," *Big Data Res.*, vol. 21, Sep. 2020, Art. no. 100145.
- [2] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven human mobility modeling: A survey and engineering guidance for mobile networking," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 1–39, Feb. 2016.
- [3] Y. Liu, L. Kong, and G. Chen, "Data-oriented mobile crowdsensing: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2849–2885, 3rd Quart., 2019.
- [4] Y. Liang, X. Zheng, and D. D. Zeng, "A survey on big data-driven digital phenotyping of mental health," *Inf. Fusion*, vol. 52, pp. 290–307, Dec. 2019.
- [5] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, Jun. 2011.
- [6] A. Cheu, A. Smith, and J. Ullman, "Manipulation attacks in local differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2021, pp. 883–900.
- [7] X. Cao, J. Jia, and N. Z. Gong, "Data poisoning attacks to local differential privacy protocols," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 1–18.
- [8] M. O. Rabin, "How to exchange secrets with oblivious transfer," *IACR Cryptol. ePrint Arch.*, vol. 2005, no. 187, 2005.
- [9] S. Taisho, T. Yuzo, and K. Youki, "Decoupling statistical trends from data volume on LDP-based spatio-temporal data collection," in *Proc. IEEE Future New. World Forum*, 2022.
- [10] F. Z. Errounda and Y. Liu, "Collective location statistics release with local differential privacy," *Future Gener. Comput. Syst.*, vol. 124, pp. 174–186, Nov. 2021.
- [11] X. Zhao, Y. Li, Y. Yuan, X. Bi, and G. Wang, "LDPart: Effective location-record data publication via local differential privacy," *IEEE Access*, vol. 7, pp. 31435–31445, 2019.
- [12] M. Kohlweiss, S. Faust, L. Fritsch, B. Gedrojc, and B. Preneel, "Efficient oblivious augmented maps: Location-based services with a payment broker," in *Proc. Int. Workshop Privacy Enhancing Technol.* Cham, Switzerland: Springer, 2007, pp. 77–94.
- [13] B. Bi, D. Huang, B. Mi, Z. Deng, and H. Pan, "Efficient LBS security-preserving based on NTRU oblivious transfer," *Wireless Pers. Commun.*, vol. 108, no. 4, pp. 2663–2674, Oct. 2019.
- [14] H. Jannati and B. Bahrak, "An oblivious transfer protocol based on ElGamal encryption for preserving location privacy," *Wireless Pers. Commun.*, vol. 97, no. 2, pp. 3113–3123, Nov. 2017.
- [15] K. C. Zeng, Y. Shu, S. Liu, Y. Dou, and Y. Yang, "A practical GPS location spoofing attack in road navigation scenario," in *Proc. 18th Int. Workshop Mobile Comput. Syst. Appl.*, Feb. 2017, pp. 85–90.
- [16] G. Qin, C. Patsakis, and M. Bourouche, "Playing hide and seek with mobile dating applications," in *Proc. IFIP Int. Inf. Secur. Conf.* Cham, Switzerland: Springer, 2014, pp. 185–196.
- [17] A. Y. Lin, K. Kuehl, J. Schoning, and B. Hecht, "Understanding 'death by GPS': a systematic study of catastrophic incidents associated with personal navigation technologies," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 1154–1166.
- [18] J. Yang, X. Cheng, S. Su, R. Chen, Q. Ren, and Y. Liu, "Collecting preference rankings under local differential privacy," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Apr. 2019, pp. 1598–1601.
- [19] J. Schwartz. (2009). *Bing Maps Tile System*. [Online]. Available: <http://msdn.microsoft.com/en-us/library/bb259689.aspx>
- [20] D. Lian, Y. Wu, Y. Ge, X. Xie, and E. Chen, "Geography-aware sequential location recommendation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2009–2019.
- [21] D. Lee and S. H. L. Liang, "Crowd-sourced carpool recommendation based on simple and efficient trajectory grouping," in *Proc. 4th ACM SIGSPATIAL Int. Workshop Comput. Transp. Sci.*, Nov. 2011, pp. 12–17.
- [22] Y. Kanemaru, S. Matsuura, M. Kakiuchi, S. Noguchi, A. Inomata, and K. Fujikawa, "Vehicle clustering algorithm for sharing information on traffic congestion," in *Proc. 13th Int. Conf. Telecommun. (ITST)*, Nov. 2013, pp. 38–43.
- [23] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
- [24] F. Charte, A. J. Rivera, M. J. D. Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowl.-Based Syst.*, vol. 89, pp. 385–397, Nov. 2015.
- [25] T. P. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Proc. Annu. Int. Cryptol. Conf.* Cham, Switzerland: Springer, 1991, pp. 129–140.
- [26] H. Lipmaa, "Verifiable homomorphic oblivious transfer and private equality test," in *Proc. Int. Conf. Theory Appl. Cryptol. Inf. Secur.* Cham, Switzerland: Springer, 2003, pp. 416–433.
- [27] J. A. Garay, P. MacKenzie, and K. Yang, "Efficient and universally composable committed oblivious transfer and applications," in *Proc. Theory Cryptogr. Conf.* Cham, Switzerland: Springer, 2004, pp. 297–316.

- [28] S. Jarecki and X. Liu, "Private mutual authentication and conditional oblivious transfer," in *Proc. Annu. Int. Cryptol. Conf.* Cham, Switzerland: Springer, 2009, pp. 90–107.
- [29] B. Bunz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell, "Bulletproofs: Short proofs for confidential transactions and more," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 315–334.
- [30] C. Bergroth, O. Järvi, H. Tenkanen, M. Manninen, and T. Toivonen, "A 24-hour population distribution dataset based on mobile phone data from Helsinki Metropolitan Area, Finland," *Sci. Data*, vol. 9, no. 1, pp. 1–19, Feb. 2022.
- [31] Z. Zong, J. Feng, K. Liu, H. Shi, and Y. Li, "DeepDPM: Dynamic population mapping via deep neural network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 1294–1301, 2019.
- [32] M. Al-Jeri, "Towards human mobility detection scheme for location-based social network," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1–7.
- [33] J. Feng, Z. Yang, F. Xu, H. Yu, M. Wang, and Y. Li, "Learning to simulate human mobility," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 3426–3433.
- [34] J. Chen, T. Pei, S. L. Shaw, F. Lu, M. Li, S. Cheng, X. Liu, and H. Zhang, "Fine-grained prediction of urban population using mobile phone location data," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 9, pp. 1770–1786, 2018.
- [35] M. Irain, J. Jorda, and Z. Mammeri, "Landmark-based data location verification in the cloud: Review of approaches and challenges," *J. Cloud Comput.*, vol. 6, no. 1, pp. 1–20, Dec. 2017.
- [36] M. Katsomallos, K. Tzompanaki, and D. Kotzinos, "Landmark privacy: Configurable differential privacy protection for time series," in *Proc. 12th ACM Conf. Data Appl. Secur. Privacy*, Apr. 2022, pp. 179–190.
- [37] M. Irain, Z. Mammeri, and J. Jorda, "Assessment of regression-based techniques for data location verification at country-level," in *Proc. 6th Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Oct. 2018, pp. 1–6.



**TAISHO SASADA** (Student Member, IEEE) received the M.Sc. degree in engineering from the Nara Institute of Science and Technology (NAIST), in 2021, where he is currently pursuing the Ph.D. degree. He is a Research Fellow (DC1) with the Japan Society for the Promotion of Science (JSPS) and a Research Assistant with NAIST. His research interests include data privacy, access control, data resampling, and machine/deep learning security.



**YUZO TAENAKA** (Member, IEEE) received the D.E. degree in information science from the Nara Institute of Science and Technology (NAIST), Japan, in 2010. He was an Assistant Professor with The University of Tokyo, Japan. He has been an Associate Professor with the Laboratory for Cyber Resilience, NAIST, since April 2018. His research interests include information networks, cybersecurity, distributed systems, and software-defined technology.



**YOUKI KADOBAYASHI** (Member, IEEE) received the Ph.D. degree in computer science from Osaka University, Japan, in 1997. Since 2013, he has been working as the Rapporteur of ITU-T Q.4/17 for cybersecurity standardization. He is currently a Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. His research interests include cybersecurity, web security, and distributed systems. He is a member of the IEEE Communications Society.

...