IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 28, 2021, accepted January 13, 2022, date of publication January 14, 2022, date of current version January 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143813

Private True Data Mining: Differential Privacy Featuring Errors to Manage Internet-of-Things Data

YUICHI SEI[1,2], **(Member, IEEE), AND AKIHIKO OHSUGA**[1], **(Member, IEEE)**
[1]Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan
[2]JST, PRESTO, Kawaguchi, Saitama 332-0012, Japan

Corresponding author: Yuichi Sei (seiuny@uec.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP18H03229, Grant JP18H03340, Grant JP18K19835, Grant JP19K12107, Grant JP19H04113, and Grant JP21H03496; and in part by the Japan Science and Technology Agency (JST), Precursory Research for Embryonic Science and Technology (PRESTO), under Grant JPMJPR1934.

ABSTRACT Available data may differ from true data in many cases due to sensing errors, especially for the Internet of Things (IoT). Although privacy-preserving data mining has been widely studied during the last decade, little attention has been paid to data values containing errors. Differential privacy, which is the de facto standard privacy metric, can be achieved by adding noise to a target value that must be protected. However, if the target value already contains errors, there is no reason to add extra noise. In this paper, a novel privacy model called true-value-based differential privacy (TDP) is proposed. This model applies traditional differential privacy to the "true value" unknown by the data owner or anonymizer but not to the "measured value" containing errors. Based on TDP, the amount of noise added by differential privacy techniques can be reduced by approximately 20% by our solution. As a result, the error of generated histograms can be reduced by 40.4% and 29.6% on average according to mean square error and Jensen–Shannon divergence, respectively. We validate this result on synthetic and five real data sets. Moreover, we proved that the privacy protection level does not decrease as long as the measurement error is not overestimated.

INDEX TERMS Data mining, data privacy, differential privacy, Internet of Things.

I. INTRODUCTION

Significant amounts of IoT data are generated every day by many different sensors, such as thermal cameras, home appliance sensors, automotive sensors, and smartphone-equipped sensors. These IoT data can be used for health monitoring [1], context-aware recommendation (or recommender) systems [2], navigation [3], and other applications.

However, sensing people or their surrounding environment might involve information that identifies an individual [4]. Thus private information is at risk of leakage. By anonymizing data based on ϵ-differential privacy [5], [6], which is the de facto standard privacy metric (ϵ represents the privacy budget), privacy leakage can be controlled. Differential privacy has been used in many studies, such as [7]–[9], as it is one of the most critical privacy metrics [10]. It is considered an important concept for data analysis [11], [12].

Local differential privacy is a specialized concept of differential privacy especially for data collection from each person. In this paper, "differential privacy," refers to "local differential privacy." A differential private value can be obtained by adding *Laplace noise* to a target value, which must be protected with respect to numerical values [5]. For categorical values, a differential private category ID can be obtained by disguising the sensed category ID with a certain probability [13], [14]. These methods are widely used to achieve ϵ-differential privacy. However, they do not consider errors in values.

In this paper, an original value with no error is referred to as a "true" value; the owner or anonymizer might not know these values. Alternatively, sensed values that might have errors are referred to as "measured" values. Existing studies regarding differential privacy do not consider true values but only measured values. Our study aims to determine whether additional noise should be added to protect privacy if the target value already contains errors. This research proposes

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu.

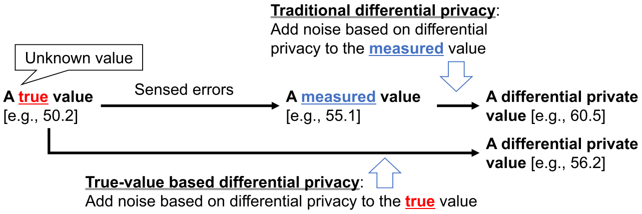

Traditional differential privacy:
Add noise based on differential privacy to the measured value
Unknown value
A true value [e.g., 50.2]
Sensed errors
A measured value [e.g., 55.1]
A differential private value [e.g., 60.5]
A differential private value [e.g., 56.2]
True-value based differential privacy:
Add noise based on differential privacy to the true value

FIGURE 1. Concept of true-value-based differential privacy (TDP). Traditional differential privacy adds differential privacy noise to the measured value. However, TDP tries to add noise to the true value.

TABLE 1. Relationship between the error distribution knowledge and the TDP.

The correctness of the knowledge	Whether we can achieve TDP
Correct	✓
Not correct (Under-estimation)	✓
Not correct (Over-estimation)	×

a new privacy model, protecting the true value as opposed to the measured value. Since the data owner might not know the true value, the true data is assumed to have a specific probability distribution, such as a normal distribution. This probability distribution is based on the data owner's or anonymizer's knowledge or the theory of errors [15], [16]. The difference between the traditional approach and the proposed true-value-based differential privacy (TDP) is described in Fig. 1. According to the TDP concept, the amount of noise added to the measured value can be reduced.

We assume that the anonymizer can estimate the distribution of measurement errors to some extent. Therefore, TDP can be achieved even if the anonymizer is incorrect, as long as they do not overestimate the amount of sensing errors. The relationship between the anonymizer's error distribution and the TDP is listed in Table 1. Therefore, if the anonymizer cannot be certain about the error distribution, they can guarantee TDP by conservatively estimating the amount of error. If the error amount is predicted to be zero, we get the same result as traditional differential privacy. Thus, TDP can reduce the amount of error introduced compared to traditional differential privacy and still achieve the desired privacy protection level specified by ϵ.

If we have no information about the error distribution, we cannot use the proposed method in this paper. However, we believe that there are many situations where it is possible to make estimates under the condition that we can underestimate the amount of error. An additional discussion on this point is given in Section VI.

In Section IV, we have conducted experiments using three synthetic datasets and three real datasets. We compared our method with the other three methods for numerical datasets and four methods for categorical datasets. The experimental results show our solution could reduce the amount of noise by approximately 20% and reduce the amount of error of generated histograms from 20% to 40% on average.

In recent years, several methods for LDP have been proposed. These LDP methods primarily estimate the distribution underlying the user data and federated learning [17]–[20]. They put noise on the values to satisfy differential privacy or randomly vary the category values. They proposed various methods to reduce the amount of noise in the differential privacy to have as little negative impact on the statistical analysis as possible. However, they do not consider sensing errors in IoT environments.

The authentication protocols for IoT are organized in detail in a survey paper by Ferrag *et al.* [21]. This survey paper categorized IoT environments into Machine-to-Machine Communications, Internet of Vehicles, Internet of Energy, and Internet of Sensors. The advantages and disadvantages of authentication protocols in each environment were summarized. One of the main objectives of an authentication protocol is to send each data to the correct entity correctly. It cannot directly solve the purpose of statistically correct analysis without sending the correct information of each data to anyone. Note that this objective is not a special objective of our study but a common objective in existing studies that collect data using differential privacy.

Badun *et al.* surveyed security and privacy issues on IoT platforms [22]. They state that most IoT platforms do not inform users about the type of information they collect and where it is shared. Therefore, some IoT platforms use new technologies that still put the data stored in the cloud under the user's control. Such techniques can protect user data; however, they do not provide a mechanism to analyze the data of many users across the board statistically.

Husnoo *et al.* organized the techniques of differential privacy in IoT environments [23]. They categorized two major usage scenarios for differential privacy: (1) a scenario where a trusted server holds the true data and only shares the statistical analysis results with third parties and (2) a scenario where data is collected on an untrusted server. Our study and several studies in literature focus on the second scenario. The disadvantage of the latter scenario is that the cumulative amount of noise added to each data becomes larger and affects the usefulness of the data [23]. Using our proposed method considering measurement errors, the cumulative amount of noise can be reduced as shown in our experimental results.

Ma *et al.* proposed an algorithm based on stochastic influence perturbation to satisfy differential privacy [24]. They assumed that a trusted central server has a whole raw dataset, and their aim was to generate a private version of the dataset. They proposed a framework for network traffic tensor data privacy protection. They used multiple-strategy differential privacy for network traffic tensor data. Their mechanism also assumed a central trusted server, although they partly used local differential privacy.

Onesimu *et al.* proposed a novel privacy protection data collection scheme for IoT-based healthcare service systems [25]. Their method uses a clustering-based anonymous model to develop an efficient privacy protection scheme that satisfies privacy requirements and protects a healthcare IoT from various privacy attacks. The proposed scheme can efficiently deal with privacy attacks such as attribute, identity, and membership disclosure; and sensitivity, similarity, and skewness attacks. However, their method does not ensure differential privacy. Thus, techniques on security and privacy in IoT environments do not often mention sensing noise. Our work is positioned as an important and pioneering effort to consider sensing noise in addressing privacy.

The motivation, the research gap, and contribution to this study are summarized below.

A. MOTIVATION

This study aims to estimate the distribution of personal data sensed in IoT environments while protecting user data by differential privacy. We assume the sensed data contains sensing noise.

B. RESEARCH GAP

Existing methods do not consider the sensing noise. Therefore, they add a lot of extra privacy noise to the sensed data.

C. CONTRIBUTION

First, we propose true-value-based differential privacy (TDP), which is a novel concept of differential privacy considering sensing noise. Second, we proposed anonymizing algorithms for numerical data and categorical data to satisfy TDP. Third, we prove that the proposed algorithms ensure TDP. Fourth, we show that the proposed algorithms can reduce the amount of differential privacy noise using synthetic and real datasets. Fifth, we show that the proposed algorithms can reduce errors in the estimated distribution of personal data using the same datasets.

The rest of the paper is structured as follows. First, the application and assumptions, along with the definition of privacy, are presented in Section II. Then, the proposed design and its mechanisms are introduced in Section III. Next, the simulation results using synthetic and five real data sets are presented in Section IV. The related methods are discussed in Section V, and several design issues of the proposed method are mentioned in Section VI. Finally, the conclusions of this work are presented in Section VII.

II. MODELS

A. APPLICATION MODEL

Currently, IoT devices can collect and estimate people's attribute information, such as location, heartbeat, health condition, age, and moving behavior. Based on these attribute data, people can use various services such as recommender systems. In addition, the data collector can also serve as a data anonymizer, anonymizing the obtained data and sending it to the data receiver (see Fig. 2).

Two kinds of attribute data are considered: The first is a numerical attribute, such as heartbeats per minute, while the second is a categorical attribute, such as a disease name (e.g., COVID-19).

The collected attribute data usually have some sensing errors since it is difficult to sense and estimate people's attributes with complete accuracy. In the worst-case scenario, some attribute data cannot be collected at all. Missing data can be estimated through multiple imputations or predictions based on regression models [26]. These estimated values exhibit a large number of errors.

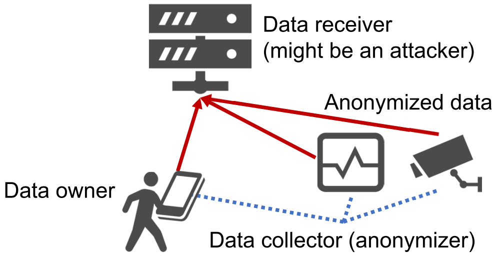


FIGURE 2. Application scenario. The data receiver, an attacker, collects user data from the data owner and data collector using privacy techniques differentially.

B. ASSUMPTIONS

Anonymizers may not know the true attribute values, but they can estimate them. However, these estimated values might contain errors. Anonymizers can estimate the error distribution of numerical attribute values. A normal distribution is considered an error model for numerical attributes since the measurement errors follow normal distributions in many cases [27]. A normal distribution is characterized by the parameter σ, which represents the standard deviation. Section VI-A contains further discussion about assuming a normal distribution. However, please note that the concept of TDP can be applied to other error models.

The wrong classification probability $p_{i \to j}$ is considered with reference to categorical attributes. This probability signifies that the true category ID is i. However, the anonymizer is unaware of the true category ID and assumes that the category ID is j.

In this paper, parameters σ and $p_{i \to j}$ for all i, j are referred to as "error parameters." Three scenarios are assumed.

1) SCENARIO I

The anonymizer knows the exact error parameters.

2) SCENARIO II

The anonymizer does not know the exact error parameters. The estimated parameters might differ from the actual parameters; however, they are not pessimistic about the degree of error. The mathematical definitions of the numerical attributes are described in Section III-A, and those of the categorical attributes are described in Section III-B.

3) SCENARIO III

The anonymizer does not know the exact error parameters and has no estimate for them.

In this paper, we do not target Scenario III. Instead, we mainly target Scenario II because Scenario I is relatively unrealistic.

C. ATTACK MODEL

The receiver of anonymized data is considered an attacker, and the attacker is considered a semi-honest entity; that is, the attacker follows the given protocol of anonymized IoT data collection. However, the attacker might try to extract individual information from each anonymized datum. Furthermore, each anonymized datum contains errors based on original sensing errors and intentionally added noise; therefore, the attacker cannot accurately estimate people's true data, but the attacker can estimate them as a particular probability distribution.

D. PRIVACY METRIC

In a privacy-preserving data mining community, differential privacy [5] is considered the most important privacy metric. Although differential privacy was originally used in a query-response database, recent studies have used it for anonymized data collection.

Suppose that a person has an attribute value and the person or anonymizer who collects the attribute value anonymizes its value. Let ϵ be a positive real number. The differential privacy is defined as follows:

Definition 1 (ϵ-Differential Privacy): Let D and D' be databases differing on one record at most. A randomized mechanism $\mathcal{A}$ satisfies ϵ-differential privacy if and only if for all $Y \subset Range(\mathcal{A})$, the following equation holds:

$$P(\mathcal{A}(D) \in Y) \leq e^{\epsilon} P(\mathcal{A}(D') \in Y) \quad \text{for all} \quad D, D'.$$

Kasiviswanathan *et al.* [28] established that this definition could be applied to anonymized data collection.

Definition 2 (Local Privacy): Let x and x' be a database of size $= 1$, and let ϵ be a positive real number. A randomized mechanism $\mathcal{A}$ satisfies ϵ-differential privacy if and only if for any output y, the following equation holds:

$$P(\mathcal{A}(x) = y) \leq e^{\epsilon} P(\mathcal{A}(x') = y) \quad \text{for all} \quad x, x'. \tag{1}$$

In this paper, it is considered that a value of x might contain sensing errors. Therefore, the focus must be placed on the true value of x, which is an unknown value, even for the data owner and the anonymizer. TDP is proposed to handle the privacy of unknown values.

Definition 3 (TDP): Let x and x' be a true value, and let ϵ be a positive real number. A measurement function $\mathcal{M}$ acquires an input x and outputs a measured value. A randomized mechanism $\mathcal{A}$ satisfies TDP if and only if for any output y, the following equation holds:

$$P(\mathcal{A}(\mathcal{M}(x)) = y) \leq e^{\epsilon} P(\mathcal{A}(\mathcal{M}(x')) = y) \quad \text{for all} \quad x, x'. \tag{2}$$

Theorem 1: In an anonymized data collection scenario, Definition 2 is the same as Definition 3 when the measured values contain no errors.

Proof: When the measured values contain no errors, the equations $x = \mathcal{M}(x)$ and $x' = \mathcal{M}(x')$ are hold. Therefore, in this case, Equations 1 and 2 are equivalent. □

TABLE 2. Notations.

ϵ	Privacy budget for differential privacy
Δ	Range of possible values for a numerical attribute
M	Number of categories for a categorical attribute
τ	Probability of estimating category ID correctly by IoT devices
σ	Standard deviation of the normal distribution
b	Scale parameter of the Laplace distribution (equal to Δ/ϵ)

III. TRUE-VALUE-BASED DIFFERENTIAL PRIVACY (TDP)

Existing studies define x and x' in Definition 2 as measured values. In this paper, they are defined as true values. The anonymization mechanisms for both numerical and categorical attributes are described.

A. NUMERICAL VALUES ANONYMIZATION

The Laplace mechanism [5], which adds noise based on the Laplace distribution can be used for numerical attributes. The theorem of the Laplace mechanism for data collection is introduced.

Theorem 2 (Laplace Mechanism): A randomized mechanism $\mathcal{A}$ realizes ϵ-differential privacy if $\mathcal{A}$ adds the Laplace noise $Lap(\Delta/\epsilon)$, where Δ is the range of possible values of the target attribute, and $Lap(b)$ returns independent Laplace random variables with the scale parameter b.

However, the Laplace mechanism does not take into account sensing errors. As a result, the noise based on the normal distribution is added to a true value as a sensing error, and additional noise based on the Laplace mechanism is added to the noisy value. This is the traditional approach, which always adds the Laplace noise, and is referred to as the **baseline approach** for numerical attributes. The resulting probability density function, representing the probability of the distance between the final noisy and true values, can be calculated by performing convolution of the normal and Laplace distribution.

Let $\mathcal{N}(x; \sigma^2)$, $\mathcal{L}(x; b)$ represent the probability density functions of the normal distribution, with the standard deviation being σ and the Laplace distribution with the scale being b. Centered distributions that peak at zero are only considered without loss of generality.

The convolution of the normal distribution with the standard deviation being σ and the Laplace distribution with the scale being b is represented by

$$\begin{aligned} \mathcal{U}(x; \sigma^2, b) &= \mathcal{N} \star \mathcal{L} = \int_{t=-\infty}^{\infty} \mathcal{N}(t; \sigma^2)\mathcal{L}(x-t; b)dt \\ &= \frac{e^{\frac{\sigma^2-2bx}{2b^2}} \left(\operatorname{erfc}\left(\frac{\sigma^2-bx}{\sqrt{2}b\sigma}\right) + e^{\frac{2x}{b}} \operatorname{erfc}\left(\frac{\sigma^2+bx}{\sqrt{2}b\sigma}\right)\right)}{4b} \end{aligned} \tag{3}$$

where erfc is the complementary error function, which is represented by

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt. \tag{4}$$

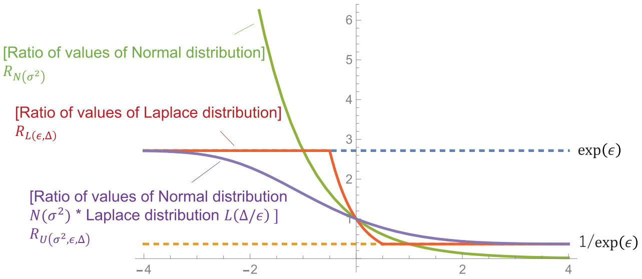


FIGURE 3. Ratio of probability density function values whose distance is Δ with respect to the normal distribution, Laplace distribution, and the convolution of the two distributions ($\sigma = \epsilon = \Delta = 1$).

It is noted that for Scenario II, the value of σ can be wrong, as long as it is not pessimistic. Let σ_t and σ represent the true standard deviation and the standard deviation the anonymizer believes, respectively. Here, pessimistic means that

$$\sigma > \sigma_t. \tag{5}$$

$\exp(\epsilon)$, $1/\exp(\epsilon)$, and the ratio of probability density function values whose distance is Δ with respect to $\mathcal{N}(x;\sigma^2)$, $\mathcal{L}(x;\Delta/\epsilon)$, and $\mathcal{U}(x;\sigma^2,\Delta/\epsilon)$, where ϵ and σ are set to one, are presented in Fig. 3. The ratio of probability density function values whose distance is Δ with respect to the normal distribution is calculated by

$$R_{\mathcal{N}(x;\sigma^2)} = \frac{\mathcal{N}(x+\Delta/2;\sigma^2)}{\mathcal{N}(x-\Delta/2;\sigma^2)} = e^{-\frac{\Delta x}{\sigma^2}}. \tag{6}$$

Equation 6, shows that $R_{\mathcal{N}(x;\sigma^2)}$ approaches ∞ when x is close to -∞. Therefore, even if σ is very large, extra noise needs to be added to achieve ϵ-differential privacy.

Similarly, in Fig. 3, $R_{\mathcal{L}(x;\epsilon,\Delta)}$ and $R_{\mathcal{U}(x;\sigma^2,\epsilon,\Delta)}$ are defined as the ratio of the probability density function values whose distance is Δ with respect to $\mathcal{L}(x;\Delta/\epsilon)$ and $\mathcal{U}(x;\sigma^2,\Delta/\epsilon)$, respectively.

The ratio of the probability density function values whose distance is Δ should exist between the lines of $\exp(\epsilon)$ and $1/\exp(\epsilon)$, according to the definition of ϵ-differential privacy. Fig. 3 shows that $R_{\mathcal{L}(x;\epsilon,\Delta)}$ and $R_{\mathcal{U}(x;\sigma^2,\epsilon,\Delta)}$ satisfy this condition; therefore, $\mathcal{L}(x;\Delta/\epsilon)$ and $\mathcal{U}(x;\sigma^2,\Delta/\epsilon)$ mechanisms achieve ϵ-differential privacy (here $\sigma = \Delta = \epsilon = 1$). Although $R_{\mathcal{U}(x;\sigma^2,\Delta/\epsilon)}$ approaches $\exp(\epsilon)$ (or $1/\exp(\epsilon)$) when $|x|$ is large, its convergence to $\exp(\epsilon)$ (or $1/\exp(\epsilon)$) is slower than that of $R_{\mathcal{L}(x;\Delta/\epsilon)}$. Consequently, the mechanism adds much more noise than required.

The algorithm proposed in this paper is simple but effective; Laplace noise is not added when the obtained Laplace noise is smaller than the predefined threshold w. Thus, the total loss is expected to become smaller (i.e., the ratio of probability density function values whose distance is Δ is expected to approach $\exp(\epsilon)$ and $1/\exp(\epsilon)$ faster).

However, the definition of an appropriate value for w is complex. If the threshold w is very large, the resulting value cannot achieve either traditional ϵ-differential privacy or TDP. Conversely, the resulting value contains unnecessary noise if the threshold w is very small.

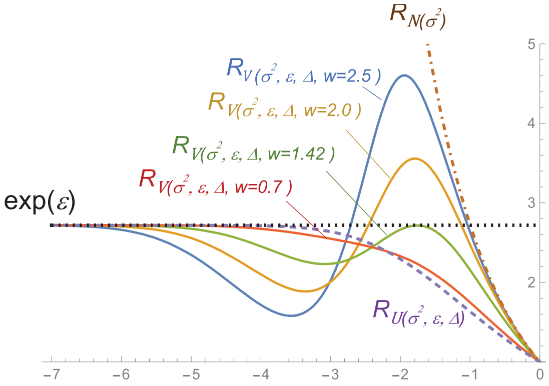


FIGURE 4. $R_{\mathcal{V}}$ for various values of w ($\sigma = \epsilon = \Delta = 1$). It can be seen that if the value of w is too large, the requirement for differential privacy is not met. Alternatively, it can be seen that if the value of w is too small, it adds more noise than necessary.

The probability density function, which adds the Laplace noise only when the noise x satisfies abs$(x) \geq w$[1] is represented by

$$\widehat{\mathcal{L}}(x;b,w) = \begin{cases} \int_{-w}^{w}\mathcal{L}(t;b)dt & x = 0 \\ \frac{e^{-x/b}}{2b} & x \geq w \\ \frac{e^{x/b}}{2b} & x \leq -w \\ 0 & otherwise. \end{cases} \tag{7}$$

Therefore, the probability density function obtained from the original sensing error and the Laplace noise defined in Equation 7 can be represented by

$$\begin{aligned} &\mathcal{V}(x;\sigma^2,b,w) \\ &= \int_{-\infty}^{\infty}\mathcal{N}(t;\sigma^2)\widehat{\mathcal{L}}(x-t;b,w)dt \\ &\quad +\mathcal{N}(x;\sigma^2)\int_{-w}^{w}\mathcal{L}(t;b)dt \\ &= \frac{e^{-\frac{w+x}{b}-\frac{x^2}{2\sigma^2}}}{4b\sigma}\times\left\{\sigma e^{\frac{1}{2}\left(\frac{2bw+\sigma^2}{b^2}+\frac{x^2}{\sigma^2}\right)}\left[\operatorname{erfc}\left(\frac{b(w-x)+\sigma^2}{\sqrt{2}b\sigma}\right)\right.\right. \\ &\quad \left.\left.+e^{\frac{2x}{b}}\operatorname{erfc}\left(\frac{b(w+x)+\sigma^2}{\sqrt{2}b\sigma}\right)\right]+2\sqrt{\frac{2}{\pi}}b\left(e^{\frac{w}{b}}-1\right)e^{\frac{x}{b}}\right\}. \end{aligned} \tag{8}$$

The ratio of probability density function values for the proposed algorithm whose distance is Δ is represented by the following:

$$R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} = \frac{\mathcal{V}(x+\Delta/2;\sigma^2,\Delta/\epsilon,w)}{\mathcal{V}(x-\Delta/2;\sigma^2,\Delta/\epsilon,w)} \tag{9}$$

The aim is to find an appropriate value of w where $R_{\mathcal{V}}$ approximates $\exp(\epsilon)$ but does not cause $R_{\mathcal{V}}$ to overestimate $\exp(\epsilon)$ or $1/\exp(\epsilon)$.

The following theorem is considered (see Fig. 4);

[1] More formally, this is a combination of a probability density function and a probability mass function.

Theorem 3: If w is near ∞, the value of $R_{\mathcal{V}}$ approaches the value of $R_{\mathcal{N}}$. If w is near zero, the value of $R_{\mathcal{V}}$ approaches the value of $R_{\mathcal{U}}$.

Proof: $\mathcal{U}(x; \sigma^2, \Delta/\epsilon)$ (Equation 3) and $\mathcal{N}(x; \sigma^2)$ (Equation 6) can be obtained by calculating the limit of $\mathcal{V}(x; \sigma^2, \Delta/\epsilon, w)$ (Equation 8) of w as w approaches zero and ∞, respectively. □

The ratio between $x + \Delta/2$ and $x - \Delta/2$ is defined in this study; therefore, the range $-w - \Delta/2 < x < 0$ can be checked considering whether or not the maximum ratio is greater than $\exp(\epsilon)$. It is noted that only the range $x < 0$ needs to be checked because $\mathcal{V}$ is symmetrical with respect to the point $(x, y) = (0, 1)$, where y represents the ratio of probability density function values whose distance is Δ.

Algorithm 1 describes the method that yields the anonymized value. In Algorithm 1, the value of w is calculated in Lines 1–16. $\mathrm{erfc}(x)$ can be computed using approximate equations, such as

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) \approx 1 - \sqrt{1 - e^{-x^2 \frac{4/\pi + 0.147x^2}{1+0.147x^2}}} \quad \text{(maximum relative error: } 1.3 \cdot 10^{-4}) \qquad (10)$$

when $x \geq 0$ from [29]. Note that we can obtain an approximate value of $\mathrm{erfc}(x)$ with $x < 0$ from the property of

$$\mathrm{erfc}(x) = 2 - \mathrm{erfc}(-x). \qquad (11)$$

After checking the approximate values, precise values need to be calculated. Mathematical tools such as Maxima,[2] which is a popular free software program, can be employed.

B. CATEGORICAL VALUES ANONYMIZATION

The randomized response mechanism [13], [14] can be used for categorical attributes. First, a sensed value is categorized into one of the predefined categories. Another category replaces that category with a certain probability, and then the resulting category ID is sent to the data receiver. The randomized response is referred to as the **baseline approach** for categorical attributes.

The retention probability of unchanging category ID is p_α and the probabilities of other IDs are $(1-p_\alpha)/(M-1)$, where M is the number of categories. The following equation

$$\max\left(\frac{p_\alpha}{(1-p_\alpha)/(M-1)}, \frac{(1-p_\alpha)/(M-1)}{p_\alpha}\right) \leq e^\epsilon \qquad (12)$$

should hold to satisfy ϵ-differential privacy. Therefore, it is set

$$p_\alpha = e^\epsilon/(M - 1 + e^\epsilon). \qquad (13)$$

Since $M \geq 2$, $p_\alpha > 0.5$ is obtained.

Let $p_{i\to j}$ represent the probability that the true category ID C_i is (mis-)classified to C_j due to sensing errors. It is assumed that the retention probability is greater than any other probabilities; that is, the following equation is assumed

$$p_{i\to i} > \max_{j\neq i} p_{i\to j}. \qquad (14)$$

It is assumed that the values of $p_{i\to j}$ for all i, j can be estimated. Let

$$\boldsymbol{p_i} = \{p_{i\to 1}, p_{i\to 2}, \ldots, p_{i\to M}\}. \qquad (15)$$

For Scenario II, these values can be wrong, as long as they are not pessimistic.

Let $p_{i\to j,t}$ and $p_{i\to j}$ represent the true probability and the probability that the anonymizer believes, respectively. Here, pessimistic estimation means that

$$\begin{cases} p_{i\to i} < p_{i\to i,t} \text{ for any } i, \\ p_{i\to j} > p_{i\to j,t} \text{ for any } i, j(i \neq j). \end{cases} \qquad (16)$$

First, the situation where the following expression is satisfied;

$$\frac{p_{i\to j}}{p_{i'\to j}} \leq e^\epsilon \text{ for all } i, i', j. \qquad (17)$$

This case holds TDP clearly. In this case, the random mechanism $\mathcal{A}$ in Definition 3 does not need to do anything. In other words, the TDP can be satisfied by outputting the input measured values as they are.

Algorithm 1 Proposed Randomization Mechanism for Numerical Attributes

```
Input: Privacy budget $\epsilon$, Standard deviation of the normal
       distribution for sensing error $\sigma$, Range of possible values
       $\Delta$, Measured value $v_s$
Output: TDP value
 1: $w_{max} \leftarrow$ sufficient large value
 2: $w_{min} \leftarrow 0$
 3: while True do
 4:   $w' \leftarrow (w_{max} + w_{min})/2$
 5:   $r \leftarrow \max_{-w-\Delta/2 \leq x \leq 0} (R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w')} - \exp(\epsilon))$.
 6:   if $r > 0$ then
 7:     $w_{max} \leftarrow w'$
 8:   else
 9:     if $w' - w_{min}$ is sufficient small then
10:       $w \leftarrow w'$
11:       Break.
12:     else
13:       $w_{min} \leftarrow w'$
14:     end if
15:   end if
16: end while
17: Generate Laplace noise $l$ based on $\mathcal{L}(0, \Delta/\epsilon)$.
18: if $l < w$ then
19:   Return $v_s$.
20: else
21:   Return $v_s + l$.
22: end if
```

[2] http://maxima.sourceforge.net/

Algorithm 2 Proposed Randomization Mechanism for Categorical Attributes

```
Input: Privacy budget $\epsilon$, Probabilities $p_{i \to j}$ for all $i, j$, measured category ID $s$, IDs of categories $K$
Output: TDP value Scenarios I and II
Calculate $p_\alpha$ from Equation 13.
if Equation 17 holds then
    Return $s$.
else
    Solve simultaneous equations 18 and obtain $\boldsymbol{x_i}$ for all $i$.
    Normalize $\boldsymbol{x_i}$ using Equation 20.
    Randomly select $j$ each with a probability $x_{i \to j}$, and return $j$.
end if
```

If Equation 17 is not satisfied, the following simultaneous equations with respect to $x_{i \to j}$ for all i and j are solved:

$$\begin{aligned} \boldsymbol{p_i} \cdot \boldsymbol{x_i} &= p_\alpha \quad \text{for} \quad i = 1, \ldots, M, \\ \boldsymbol{p_i} \cdot \boldsymbol{x_j} &= \frac{1 - p_\alpha}{M - 1} \quad \text{for} \quad i, j = 1, \ldots, M \quad \text{s.t.} \quad i \neq j, \end{aligned} \tag{18}$$

where

$$\boldsymbol{x_i} = \{x_{1 \to i}, x_{2 \to i}, \ldots, x_{M \to i}\} \tag{19}$$

and $\cdot$ represents the scalar product of two vectors.

The value of $x_{i \to i}$ could be greater than one and the value of $x_{i \to j}$ could be less than zero. Therefore, the obtained values are normalized by

$$\begin{aligned} x_{i \to i} &\leftarrow \min(1, x_{i \to i}) \quad \text{for} \quad i = 1, \ldots, M, \\ x_{i \to j} &\leftarrow \max(0, x_{i \to j}) \quad \text{for} \quad i, j = 1, \ldots, M \quad \text{s.t.} \quad i \neq j. \end{aligned} \tag{20}$$

Finally, when the measured category ID is C_i, the anonymizer generates the anonymized version C_j with probability $x_{i \to j}$.

Algorithm 2 shows the method which yields the anonymized category ID.

C. PROOF OF ACHIEVING TRUE VALUE-BASED DIFFERENTIAL PRIVACY

Next, it is proved that the proposed algorithms (for Scenarios I and II) realize TDP.

1) NUMERICAL ATTRIBUTES

Initially, Scenario I is considered. Since Algorithm 1 ensures that $1/\exp(\epsilon) \leq R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} \leq \exp(\epsilon)$ for the true value if σ is correct, it is able to achieve TDP based on Definition 3.

Next, Scenario II is considered. It is assumed that the anonymizer's knowledge about sensing errors is not correct, but their knowledge about measurement errors is not pessimistic. The concept "pessimistic" is defined in Equation 5 regarding numerical attributes.

Let the ratio of the probability density function values whose distance is Δ with respect to $\mathcal{N}(x; \sigma^2)$ be $R_{\mathcal{N}(x;\sigma^2)}$. By differentiating $R_{\mathcal{N}(x;\sigma^2)}$ with respect to σ, it is obtained

$$\frac{\partial R_{\mathcal{N}(x;\sigma^2)}}{\partial \sigma} = \frac{2\Delta e^{-\frac{\Delta x}{\sigma^2}} x}{\sigma^3}. \tag{21}$$

When x is less than zero, the value of the differentiation of $R_{\mathcal{N}(x;\sigma^2)}$ with respect to σ is always less than zero. Therefore, if σ becomes larger, the value of $R_{\mathcal{N}(x;\sigma^2)}$ becomes smaller. It can be concluded that $R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)}$ becomes smaller when σ becomes larger, since the proposed probability density function $\mathcal{V}(x; \sigma^2, \Delta\epsilon, w)$ is a convolutional function of $\mathcal{N}(x; \sigma^2)$ and Equation 7, which does not depend on σ. Therefore, if the knowledge about measurement errors of the anonymizer is not pessimistic, then $R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} \leq R_{\mathcal{V}(x;\sigma_t^2,\epsilon,\Delta,w)}$ for $x \leq 0$. If the anonymizer sets the value of error parameters to pessimistic (i.e., set σ to a small value), the amount of noise added by the proposed mechanism is larger than the necessary amount. Although the usefulness of the proposed algorithm becomes worse in this case, the ratio of the anonymization probabilities generated by the proposed mechanism from two neighboring databases is within the range between $\exp(\epsilon)$ and $1/\exp(\epsilon)$, with some extra space available. However, the total loss of the proposed mechanism is less than the baseline approach, even in this case. When $x > 0$ is considered, the discussion is similar, and then $R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} > R_{\mathcal{V}(x;\sigma_t^2,\epsilon,\Delta,w)}$ for $x > 0$.

Since $1/\exp(\epsilon) \leq R_{\mathcal{V}(x;\sigma_t^2,\epsilon,\Delta,w)} \leq \exp(\epsilon)$ for σ_t^2, then $1/\exp(\epsilon) \leq R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} \leq \exp(\epsilon)$ for σ^2. Therefore, Definition 3 holds.

2) CATEGORICAL ATTRIBUTES

Initially, Scenario I is considered. It is assumed that the attacker obtains a category ID γ as the anonymized version of a categorical attribute. Let $P(v_a = \gamma | v_t = i)$ represent the anonymized version of the category ID is γ when the probability that the true category ID is i. The proposed mechanism ensures that

$$P(v_a = \gamma | v_t = i) = \begin{cases} \dfrac{e^\epsilon}{M - 1 + e^\epsilon} & (i = \gamma) \\ \dfrac{1 - \frac{e^\epsilon}{M-1+e^\epsilon}}{M - 1} & (\mathit{otherwise.}) \end{cases} \tag{22}$$

when we ignore the process of Equation 20. The ratio of two equations of Equation 22 is e^ϵ or $1/e^\epsilon$. Therefore, Definition 3 hold. Based on the post-processing property of differential privacy, the resulted values of the process of Equation 20 also satisfies TDP.

Next, Scenario II is considered. It is assumed that the anonymizer's knowledge about sensing errors is not correct but their knowledge about measurement errors is not pessimistic. Let $x_{i \to j,t}$ and $x_{i \to j}$ represent the disguising probabilities based on the true error parameters and the believed error parameters, respectively. If the error parameters are not pessimistic, then

$$\begin{cases} x_{i\to j,t} \geq x_{i\to j} & (i=j) \\ x_{i\to j,t} \leq x_{i\to j} & (otherwise.) \end{cases} \tag{23}$$

Therefore

$$\begin{cases} P(v_a=\gamma|v_t=i) \leq \dfrac{e^{\epsilon}}{M-1+e^{\epsilon}} & (i=\gamma) \\ P(v_a=\gamma|v_t=i) \geq \dfrac{1-\frac{e^{\epsilon}}{M-1+e^{\epsilon}}}{M-1} & (otherwise.) \end{cases} \tag{24}$$

From Equations 14 and 24, it is concluded that Definitions 3 hold.

D. ANALYSIS

1) NUMERICAL ATTRIBUTES

The proposed mechanism avoids the addition of Laplace noise if the generated Laplace noise l is less than threshold w. Then, the avoidance (or skipping) ratio can be calculated by

$$\int_{-w}^{w} \mathcal{L}(x;\Delta/\epsilon)dx = 1-e^{-\epsilon w/\Delta}. \tag{25}$$

Let $\eta_{\mathcal{U}}$ and $\eta_{\mathcal{V}}$ represent the expected values of Laplace noise addition with respect to the baseline approach and the proposed mechanism, respectively. The value of $\eta_{\mathcal{U}}$ can be calculated by

$$\eta_{\mathcal{U}} = \int_{-\infty}^{\infty} |x|\cdot\mathcal{L}(x;\Delta/\epsilon)dx = \frac{\Delta}{\epsilon}, \tag{26}$$

and the value of $\eta_{\mathcal{V}}$ can be calculated by

$$\begin{aligned} \eta_{\mathcal{V}} &= \int_{-\infty}^{-w} -x\mathcal{L}(x;\Delta/\epsilon)dx + \int_{w}^{\infty} x\mathcal{L}(x;\Delta/\epsilon)dx \\ &= e^{-\frac{w\epsilon}{\Delta}}(\frac{\Delta}{\epsilon}+w) \end{aligned} \tag{27}$$

Theorem 4: $R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)}$ represents the ratio of probability density function values whose distance is Δ for the proposed mechanism. It approaches $\exp(\epsilon)$ if x is close to $-\infty$, and approaches $1/\exp(\epsilon)$ if x is close to ∞; that is

$$\lim_{x\to-\infty} R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} = e^{\epsilon} \tag{28}$$

$$\lim_{x\to\infty} R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} = 1/e^{\epsilon} \tag{29}$$

Proof: $R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)}$ can be represented by

$$\frac{\gamma e^{-\frac{(\Delta+2x)^2}{8\sigma^2}} + \epsilon e^{\frac{\epsilon(\epsilon\sigma^2-\Delta^2-2\Delta x)}{2\Delta^2}} \frac{\mathrm{erfc}\left[\frac{\alpha-x}{\sqrt{2}\sigma}\right]+e^{\frac{2\epsilon x}{\Delta}+\epsilon}\mathrm{erfc}\left[\frac{\beta+x}{\sqrt{2}\sigma}\right]}{4\Delta}}{\gamma e^{-\frac{(\Delta-2x)^2}{8\sigma^2}} + \epsilon e^{\frac{\epsilon(\epsilon\sigma^2+\Delta^2-2\Delta x)}{2\Delta^2}} \frac{\mathrm{erfc}\left[\frac{\beta-x}{\sqrt{2}\sigma}\right]+e^{\frac{2\epsilon x}{\Delta}-\epsilon}\mathrm{erfc}\left[\frac{\alpha+x}{\sqrt{2}\sigma}\right]}{4\Delta}},$$

where

$$\begin{aligned} \alpha &= \frac{\epsilon\sigma^2}{\Delta}+w-\frac{\Delta}{2}, \quad \beta = \frac{\epsilon\sigma^2}{\Delta}+w+\frac{\Delta}{2}, \\ \gamma &= \frac{1-e^{-\frac{\epsilon w}{\Delta}}}{\sqrt{2\pi}\sigma}. \end{aligned} \tag{30}$$

Since the convergence of erfc$[x]$ with $x \to \infty$ to zero is more rapid than that of $\exp(x)$ with $x \to \infty$ and since

$$\lim_{x\to-\infty} \mathrm{erfc}[x] = 2, \quad \lim_{x\to\infty} \mathrm{erfc}[x] = 0, \tag{31}$$

it is obtained

$$\lim_{x\to-\infty} R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} = \lim_{x\to-\infty} e^{-\epsilon}\frac{e^{\frac{2\epsilon x}{\Delta}+\epsilon}}{e^{\frac{2\epsilon x}{\Delta}-\epsilon}} = e^{\epsilon}. \tag{32}$$

Next, $x \to \infty$ is considered. Using l'Hopital's rule, it is obtained

$$\lim_{x\to\infty} e^{x}\mathrm{erfc}[x] = \lim_{x\to\infty}\frac{\mathrm{erfc}[x]}{e^{-x}} = \lim_{x\to\infty}\frac{\frac{2e^{-x^2}}{\sqrt{\pi}}}{e^{-x}} = 0, \tag{33}$$

since the differentiation of erfc$[x]$ yields $-\frac{2e^{-x^2}}{\sqrt{\pi}}$.

From Equations 30, 31, and 33, it is obtained

$$\lim_{x\to\infty} R_{\mathcal{V}(x;\sigma^2,\epsilon,\Delta,w)} = e^{-\epsilon}. \tag{34}$$

□

2) CATEGORICAL ATTRIBUTES

Let $\zeta_{\mathcal{U}}$ and $\zeta_{\mathcal{V}}$ represent the probabilities that the true category ID is equivalent to the anonymized category ID corresponding to the baseline approach and the proposed mechanism, respectively. The baseline approach represents a method that always adds the Laplace noise with respect to numerical attributes and the randomized response method with respect to categorical attributes, as described in Sections III-A and III-B. Assuming that the true category ID is i,

$$\zeta_{\mathcal{U}} = p_{i\to i}\cdot p_{\alpha} + \sum_{j} p_{i\to j}\cdot\frac{1-p_{\alpha}}{M-1}, \tag{35}$$

and

$$\zeta_{\mathcal{V}} = p_{i\to i}\cdot x_{i\to i} + \sum_{j} p_{i\to j}\cdot x_{j\to i}. \tag{36}$$

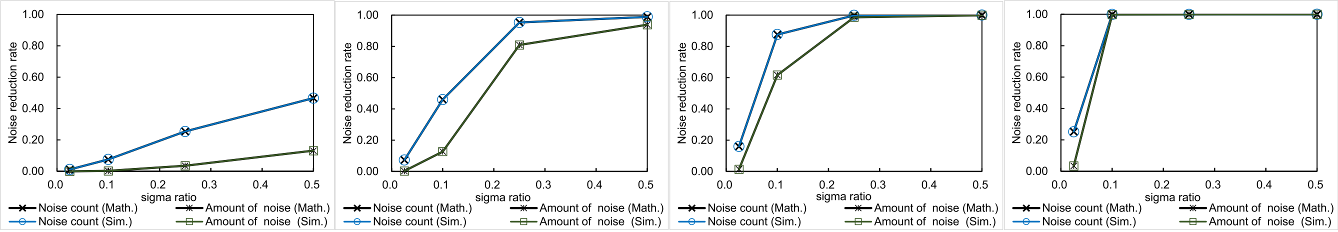
IV. EVALUATION

A. PARAMETERS SETTING

The value of ϵ and error parameters σ and $\boldsymbol{p_i}$ for all i need to be set to realistic values.

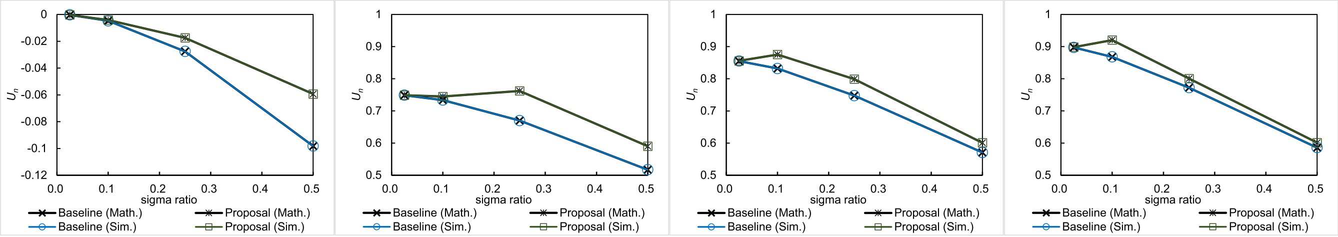
1) VALUE OF ϵ

Apple's deployment ensures that ϵ is equal to 1 or 2 per each datum [30]. An Apple's differential privacy team set $\epsilon = 2, 4, 8$ for their evaluations [31]. In the paper that proposed RAPPOR [32], which was developed by Google, $\epsilon = \log(3)$ is used as the main setting. Based on these settings, ϵ is set in the range 1–10. For this range, when Δ is equal to 100, the absolute value of the average noise added by the Laplace mechanism is in the range 5–50. In this case, privacy can be considered to be sufficiently protected. It is noted that if Δ is multiplied by a, the average noise is also multiplied by a. For categorical attributes, when the number of categories M is 2, the retention probability value ranges between 73.11%



(a) $\epsilon = 1$. (b) $\epsilon = 4$. (c) $\epsilon = 7$. (d) $\epsilon = 10$.

FIGURE 5. Reduction rate of the proposed mechanism with respect to noise addition counts and amount of Laplace noise. (The results are with $\Delta = 10$. Results with $\Delta = 100$ and $\Delta = 1000$ are almost the same.)



(a) $\epsilon = 1$. (b) $\epsilon = 4$. (c) $\epsilon = 7$. (d) $\epsilon = 10$.

FIGURE 6. U_n results. (The results are with $\Delta = 10$. Results with $\Delta = 100$ and $\Delta = 1000$ are almost the same.)

and 99.996% if ϵ is set in the range 1–10. A value of 10 for ϵ means that when the value of M is small, we have a situation where the privacy protection level is very low. Therefore, for categorical attributes, performance evaluation at small values of ϵ is especially important.

2) VALUE OF σ

When the standard deviation is σ, the average sensing error $ASE(\sigma)$ is described as

$$ASE(\sigma) = \int_{x=-\infty}^{\infty} |x| \cdot \mathcal{N}(x; \sigma^2)dx = \sqrt{\frac{\pi}{2}}\sigma. \tag{37}$$

When σ is set to 1/40 of the value of Δ (which is the range of possible values), $ASE(\sigma)$ is 2.0% of the value of Δ. In this case, the IoT device sensing the attribute value is considered to have high accuracy. Similarly, when σ is set to 1/10, 1/4, and 1/2 of the value of Δ, respectively, the values of $ASE(\sigma)$ become 8.0% (relatively high accuracy), 20% (relatively low accuracy), and 40% (low accuracy) of the value of Δ, respectively. Based on this analysis, σ is set in the range 1/40 to 1/2 of the value of Δ.

3) VALUE OF p_i

$p_{i\to i}$ for all i is set to the same value, which is referred as τ. τ is set in the range 0.3–0.9. This means that the IoT device is able to sense a person's attributes and that it can correctly judge the attribute category with a probability value from 0.3 (low accuracy) to 0.9 (high accuracy). $p_{i\to j}$ for all $i, j (i \neq j)$ is set to another value, that is,

$$p_{i\to j} = \frac{1-\tau}{M-1}. \tag{38}$$

B. UTILITY METRIC

The data receiver aims to use the anonymized value for several services. Therefore, the estimated value should be close to the true value. Let N represent the number of people whose attribute values are collected. Let v_i and $\widetilde{v_i}$ represent the true value of person i and an anonymized one, respectively.

The utility is defined below with respect to numerical attributes:

$$U_n = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|v_i - \widetilde{v_i}|}{\Delta}\right), \tag{39}$$

While the utility is defined as follows with respect to categorical attributes:

$$U_c = \frac{1}{N}\sum_{i=1}^{N}\delta_{v_i,\widetilde{v_i}}, \tag{40}$$

where $\delta_{i,j}$ is the Kronecker delta

$$\delta_{i,j} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j). \end{cases} \tag{41}$$

Both metrics are considered superior if their values are significant.

Some methods can estimate statistical values (e.g., averages) or generate cross-tabulation from the collected data. If the aim is to generate cross-tabulation, a total loss, which compares the true cross-tabulation with the generated cross-tabulation, should be used. However, in this paper, the aim is mainly focused on individual data; that is, the aim is not to do a statistical analysis but to use *each person* 's attribute value because IoT-related services, such as health monitoring, context-aware recommender systems, and navigation

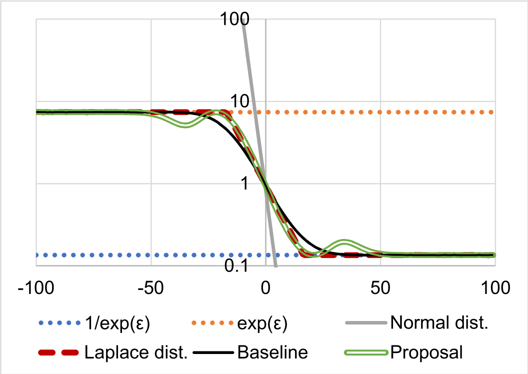


FIGURE 7. Example of simulation results: the ratio of probability distributions for numerical attributes ($\epsilon = 2$, $\Delta = 100$, $\sigma = 25$).

described in Section I need to analyze an individual's attribute value.

C. NUMERICAL VALUE RESULTS

Δ is set in the range 10–1,000, ϵ is set in the range 1–10, and σ is set in the range 1/40 of the value of Δ to 1/2 of the value of Δ. The number of times the proposed mechanism avoided the addition of Laplace noise to a measured value was evaluated. It was also evaluated how the proposed mechanism was able to reduce the average amount of Laplace noise. Results with Δ being equal to 10 are shown in Fig. 5. Computed results based on Equations 25, 26, and 27 are also presented in Fig. 5. Results with Δ =100 and Δ =1000 are almost the same as those in Fig. 5; therefore, they are not shown.

Computed results based on Equations 25, 26, and 27 are in close agreement with simulation results, for all parameter settings. The proposed mechanism reduced the number of times Laplace noise is added and the corresponding average Laplace noise. Large values of σ or ϵ result in a large reduction rate. A large value of σ means that a large sensing error noise is already added to a true value, while a large value of ϵ means that the privacy protection level is not high; that is, a large amount of noise is not needed. Therefore, the proposed mechanism can reduce additional Laplace noise, especially when the values of σ and ϵ are large. From Equation 6, it is concluded that the addition of noise cannot be completely avoided. However, Fig. 5 indicates that the noise skipping ratio approaches one.

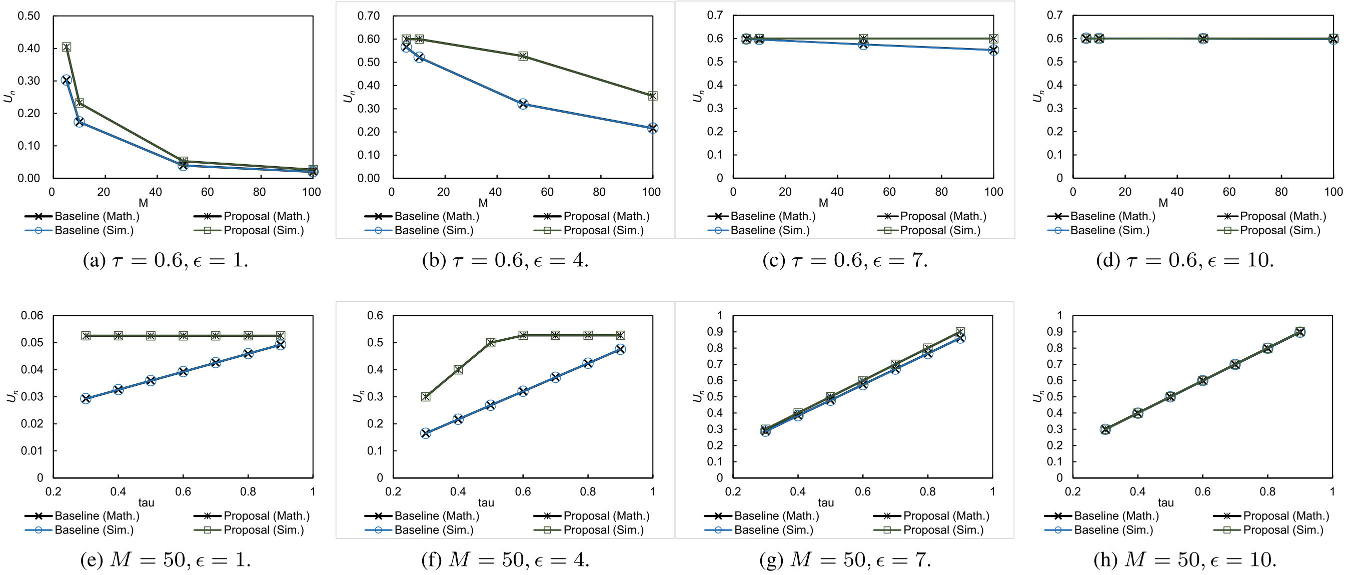
U_n was evaluated using Equation 39 using the same values for Δ and ϵ as above (Fig. 6). Since a large σ results in a low U_n (i.e., high total loss), even if none of the privacy protection mechanisms are conducted, the difference between the proposed mechanism and the baseline approach is small. This is true not only when the value of σ is minimal but also when the value of σ is substantial. However, if σ is set to a medium value, the proposed mechanism can reduce the total loss U_n by 25%–40% compared with the baseline approach. When ϵ is set to one, the difference between the proposed mechanism and the baseline approach is minor. However, when the value of ϵ is equal to one, the average absolute value of the Laplace noise to be added is about 50 when Δ is equal to 100. This amount of noise seems to be very large. Therefore, in the usual case, the value of ϵ should be larger.

The actual ratio of probability density function values whose distance is Δ was determined by conducting simulations. True values that should be protected were set to $-\Delta/2$ and $\Delta/2$. Normal distribution's noise was randomly added to the true values independently. The noise-added values were anonymized by the proposed mechanism and the baseline approach, respectively. Histograms were created for the range -3Δ - 3Δ. The number of bins was 200. This simulation was repeated 2^{31} times. In Fig. 7, an example of the average result with $\epsilon = 2$, $\Delta = 100$, and $\sigma = 25$ is presented. The ratio of the probability density function values of the normal distribution and the Laplace distribution, along with $\exp(\epsilon)$ and $1/\exp(\epsilon)$ functions, are also shown as a reference. The results for both the proposed and the baseline approach exist within the range $\exp(\epsilon)$ - $1/\exp(\epsilon)$. Therefore, it is concluded that both mechanisms (for Scenarios I and II) achieve TDP. The ratio of the probability density function values of the Laplace distribution is the same as $\exp(\epsilon)$ and $1/\exp(\epsilon)$ in the range $x < -\Delta/2$ and $\Delta/2 < x$; therefore, the Laplace mechanism is the best if the measured values have no errors. Regarding the proposed mechanism, the ratio of the probability density function values reaches $\exp(\epsilon)$ and $1/\exp(\epsilon)$ at about $x = -\Delta/2$ and $x = \Delta/2$. However, this ratio is a little far from $\exp(\epsilon)$ and $1/\exp(\epsilon)$ at $x = -40$ and $x = 40$. On the contrary, the ratio of the probability density function values in the baseline approach reaches $\exp(\epsilon)$ and $1/\exp(\epsilon)$ at about $x = -30$ and $x = 30$. Note that the values of the probability density functions are large when x is near zero; therefore, a high utility can be achieved if the ratio is near $\exp(\epsilon)$ and $1/\exp(\epsilon)$ when x is near zero. Hence, the proposed mechanism can achieve high utility (i.e., low total loss) compared with the baseline approach.

Additional simulations were conducted with other parameter settings. As a result, it was confirmed that the ratio of the probability density function values of the proposed mechanism exists within the range $\exp(\epsilon)$ - $1/\exp(\epsilon)$, except for those results with considerable variation due to the number of samples in each bin being too small.

D. CATEGORICAL VALUE RESULTS

The value of ϵ was set in the range 1–10, the value of M was set in the range 5–100, and the value of τ was set in the range 0.3–0.9. A true category ID was set to a random integer, and the category ID with probability $1 - \tau$ was changed. Then, category ID was randomized by the baseline mechanism and by the proposed mechanism. This simulation was repeated for 2^{31} times. Results with ϵ being equal to one are shown in Fig. 8. Simulation results along with computed results calculated from Equations 35 and 36 are also presented. A close agreement is observed between simulated and computed results.



(a) $\tau = 0.6, \epsilon = 1.$
(b) $\tau = 0.6, \epsilon = 4.$
(c) $\tau = 0.6, \epsilon = 7.$
(d) $\tau = 0.6, \epsilon = 10.$
(e) $M = 50, \epsilon = 1.$
(f) $M = 50, \epsilon = 4.$
(g) $M = 50, \epsilon = 7.$
(h) $M = 50, \epsilon = 10.$

FIGURE 8. U_c **results.**

The values of U_c obtained by the proposed method are larger than or equal to those obtained using the baseline approach for all parameter settings. When M is large or ϵ is small, the values of U_c are small for both mechanisms since it is difficult to maintain high accuracy for both mechanisms in such cases. However, in other cases, the proposed mechanism reduces the total loss compared with that of the baseline approach, especially when ϵ is small, i.e., the privacy protection level is high. When ϵ is large, the experimental results of the proposed method are similar to those of other methods. The value of ϵ is large enough so that the noise added to achieve differential privacy is very small. This is why there was no difference in accuracy between the methods. Therefore, it is more important to experiment when the value of ϵ is small.

Next, a true category ID is set to 1, and M is set to 10. The number of times each category ID was selected as a randomized category ID was counted. Let c_{max} and c_{min} represent the number of maximum times and the number of minimum times, respectively. Simulation results for the ratio c_{max}/c_{min} are shown in Fig. 9. exp(ϵ) is also shown as a reference. Since the results of both the proposed method and the baseline approach are equal to or less than exp(ϵ), it is concluded that both mechanisms achieve TDP for true data in Scenarios I and II. Compared with the proposed mechanism, the result based on the baseline approach is far from the exp(ϵ) line; therefore, it is concluded that the proposed mechanism is capable of achieving high utility (i.e., low total loss).

E. REAL DATA SET RESULTS

Simulations were conducted using a real data set called the Adult data set [33], which is a widely used benchmark in the research area of privacy-preserving data mining. It consists of six numerical attributes and nine categorical attributes and has 30,162 records after eliminating unknown values.

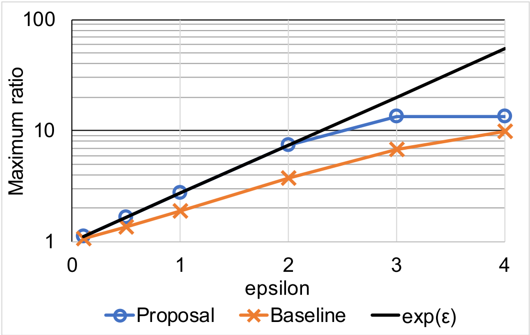


FIGURE 9. Simulation results: the ratio of probability distributions for categorical attributes ($M = 10$, $\tau = 0.6$). Since both Proposal and Baseline are smaller than the value of exp(ϵ), they both satisfy the requirement of differential privacy. Furthermore, since the Proposal is closer to the value of exp(ϵ) than Baseline, it can put more appropriate noise in terms of data utility.

It was assumed that each value of the Adult data set was true. It was also assumed that IoT devices estimated age, sex, race, and native country. using estimation methods [38]–[40]. For numerical attributes, σ was set to 0.1 of the value of Δ, and ϵ was set to 8. For categorical attributes, τ was set to 0.6, and ϵ was set to 2.

Simulation results are shown in Table 3. The names of the attributes along with Δ and M are also shown. The proposed mechanism was able to increase U_n to approximately 92% from approximately 85% for all numerical attributes and increase U_c by a maximum of 20% for categorical attributes compared with the baseline approach. These results showed that the proposed mechanism could increase the utility (i.e., reduce total loss) for real data sets.

TABLE 3. **Adult data set results [33].**

(a) U_n Results: Numerical attributes.

Attribute name	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
Δ	73	1470936	15	99999	4356	98
Baseline	0.8472	0.8474	0.8472	0.8474	0.8472	0.8473
Proposal	**0.9171**	**0.9172**	**0.9172**	**0.9172**	**0.9172**	**0.9172**

(b) U_c Results: Categorical attributes.

Attribute name	workclass	education	marital-status	occupation	relationship	race	sex	native-country	salary
M	7	16	7	14	6	5	2	41	2
Baseline	0.3610	0.2162	0.3612	0.2367	0.3901	0.4245	0.5766	0.1020	0.5760
Proposal	**0.5514**	**0.3305**	**0.5520**	**0.3622**	**0.5961**	**0.5997**	**0.5999**	**0.1559**	**0.5996**

TABLE 4. **Four real data sets results.**

(a) U_n results: Numerical attributes of AReM data set [34].

Attribute name	avg_rss12	var_rss12	avg_rss13	var_rss13	avg_rss23	var_rss23
Δ	56.25	17.24	35	11.42	40.33	13.61
Baseline	0.8472	0.8473	0.8472	0.8474	0.8474	0.8472
Proposal	**0.9182**	**0.9182**	**0.9181**	**0.9181**	**0.9182**	**0.9182**

(b) Results of U_c: Categorical attributes of three data sets.

Data set name	ADL data set [35]	RFID data set [36]	Localization data set [37]
M	10	4	11
Baseline	0.2957	0.4651	0.2780
Proposal	**0.4527**	**0.5999**	**0.4248**

Finally, simulations were conducted using other real data sets with the same parameter settings as above.

A data set of activities based on multisensor data fusion (AReM data set) [34] was used corresponding to the numerical attributes. The set consists of 42,239 instances of six numerical attributes.

A data set of daily living activities recognition using binary sensors (ADL data set) [35], a data set of healthy older people activities using a non-battery wearable sensor (RFID data set) [36], and a data set of localization for people's activity (Localization data set) [37] were used corresponding to categorical attributes. The numbers of instances are 741, 75,128, and 164,860, respectively.

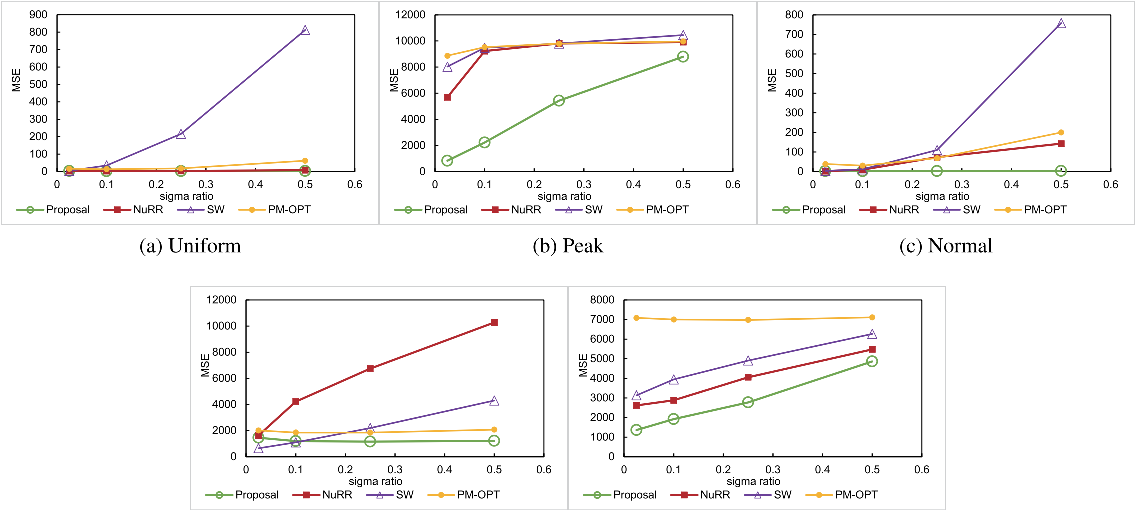
Simulation results are shown in Table 4. These results showed that the proposed mechanism outperforms the baseline approach for all data sets used in this study.

F. HISTOGRAM GENERATION

Several studies on local differential privacy have been conducted to generate accurate histograms of attribute values. In this section, we compare the accuracy of histograms generated from our proposed method and that from state-of-the-art methods. Li *et al.* proposed the square wave (SW) method for numerical attributes [19], which uses the expectation-maximization algorithm, and repeats the E-step and M-step many times. In this paper, we set the number of these iterations as 100,000. Gu *et al.* [18] proposed IDUE based on Google's RAPPOR [32] (IDUE(R)) and IDUE based on OUE [41] (IDUE(O)) for categorical attributes. Sei and Ohsuga [42] proposed an algorithm for both numerical and categorical attributes, referred to as the NuRR method. Murakami and Kawamoto proposed a utility-optimized RAPPOR (uRAP) technique for categorical characteristics [17]. uRAP assumes that non-sensitive data exist in personal data and does not protect them. However, it ensures differential privacy for sensitive data and can realize high utility. This paper, like most prior studies, assumes that all data should be protected by differential privacy; nonetheless, uRAP can be used in such situations. Zhao *et al.* proposed several strategies for differentially private data collection for numerical attributes [20]. For generating histograms, PM-SUB and PM-OPT can be used. Since PM-SUB is a simple version of PM-OPT, PM-OPT is used for this evaluation. Zhao *et al.* also proposed a Three-Output mechanism with only three discrete output possibilities. For example, regardless of an input value, Three-Output outputs $-C$, 0, or C where the value of C is determined based on ϵ. Therefore, generating an accurate histogram is challenging, although the performance of Three-Output is very high to obtain average values from differentially private data.

In summary, IDUE(R), IDUE(O), NuRR, and uRAP were compared with our proposed method for categorical attributes; SW, NuRR, and PM-OPT were compared with our proposed method for numerical attributes.

In details, we measured the mean square error (MSE) of the difference between an original histogram and that generated from anonymized values. We generated a histogram per attribute for the evaluation. Note that generating a histogram of multidimensional attributes can easily be achieved by targeting the power set of attribute



(a) Uniform

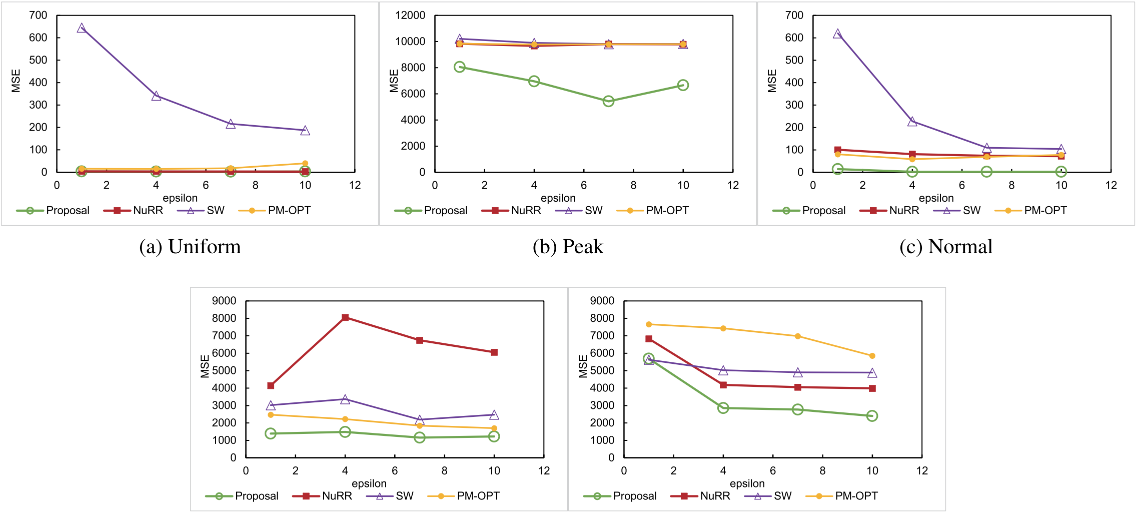
(b) Peak

(c) Normal

(d) AReM

(e) Adult

FIGURE 10. MSEs of generated histograms on numerical datasets with varying the sigma ratio.



(a) Uniform

(b) Peak

(c) Normal

(d) AReM

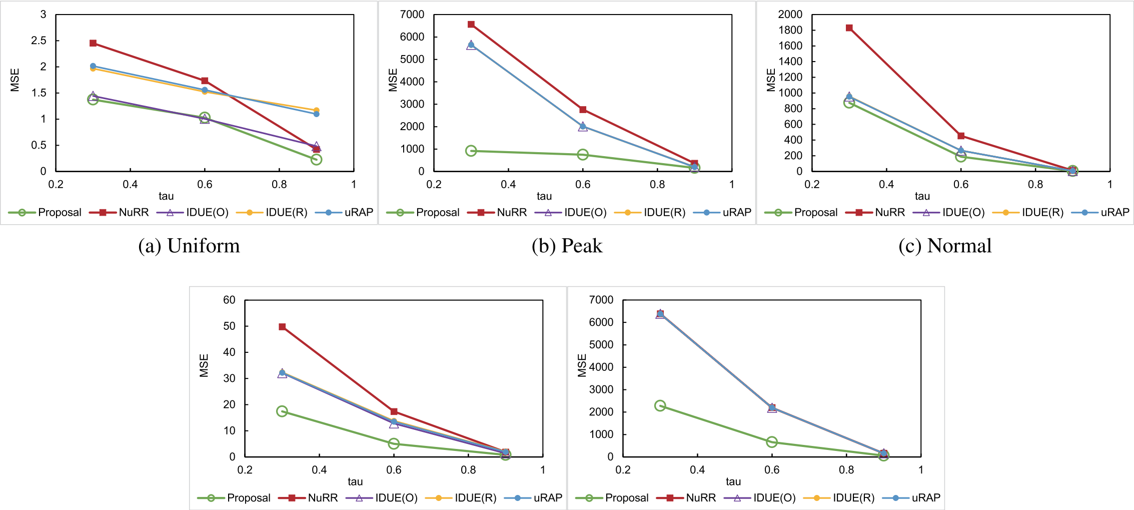
(e) Adult

FIGURE 11. MSEs of generated histograms on numerical datasets with varying ϵ.

values. Real datasets have multiple attributes, therefore, the MSE is calculated by averaging the MSE of each attribute.

First, we conducted experiments on numerical attributes. The number of bins of the histogram was set to 100. Three synthetic datasets (uniform, peak, and normal) and two real datasets (AReM and Adult) were used. As for synthetic datasets, the number of records was set to 10,000 and Δ was set to 1.0. In the uniform dataset, the value of each record was randomly generated in [0,1]. The values of all records were set to 0.5 in the peak dataset. For the normal dataset, each value was sampled as an independent and identically distributed random variable in a normal distribution with a mean of 0.5. The value of ϵ varied from 1 to 10, and the sigma ratio varied within the set $\{0.025, 0.1, 0.25, 0.5\}$. The default values were set to 7 and 0.25 for ϵ and sigma ratio, respectively. The results of varying the sigma ratio and ϵ are shown in Figs. 10 and 11, respectively.

In general, MSE for all methods in the peak dataset is the largest compared to other datasets since MSE calculates the squared value of the difference between the original and estimated histograms. Thus, the larger the difference between the values of the bins of the original histogram, the larger the MSE. The bin value corresponding to 0.5 in the peak dataset is 10,000. Suppose the estimated value is 9,000, the MSE is $1000^2/100 = 10,000$. However, for the uniform dataset, the true value of each bin in the original histogram is 100. If the estimated value for each bin is 90, then the MSE is



(a) Uniform

(b) Peak

(c) Normal

(d) ADL

(e) Adult

FIGURE 12. MSEs of generated histograms on categorical datasets with varying τ.

$100 * 10^2/100 = 100$. Thus, the more variation there is in the distribution of values, the larger the value of MSE tends to be.

The larger the sigma ratio, the larger the MSE, which is a natural result since a large sigma ratio means a large observation error (Fig. 10). The MSEs of the proposed method do not vary much on the AReM dataset. Although this result is unexpected, we consider that the reduction of observation noise has been successful.

The larger the ϵ, the smaller the MSEs (Fig. 11). This is because, in the proposed and NuRR methods, larger ϵ implies smaller Laplace noise. Similarly, for SW, the larger the ϵ, the smaller the difference between the privatized and original values.

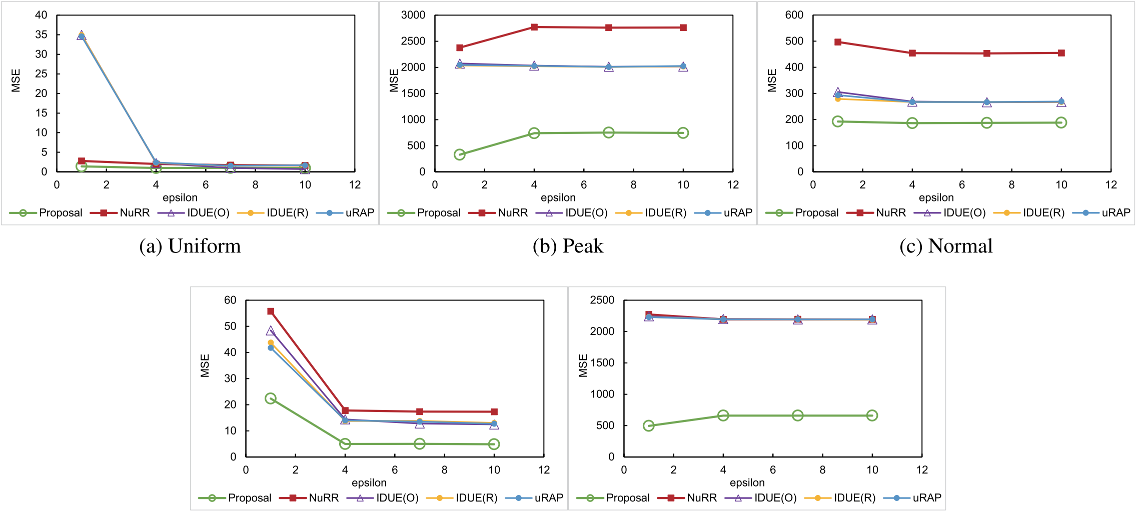
Results show that the MSE of the proposed method is the smallest in all conditions and datasets. MSEs for SW are large because SW does not consider observation noises. However, with small observation noise (i.e., small sigma ratios), the MSEs of SW can be smaller than that of the proposed method (in the AReM dataset). Authors in [19] highlighted the accuracy of SW depends on the characteristics of the datasets. The accuracy of SW can be worse if the dataset has large spikes in the distribution. The AReM dataset has larger spikes than the Adult dataset.

Next, we conducted experiments using categorized datasets. The synthesized numerical datasets were converted into categorical datasets with 10 categories. Moreover, we used the ADL and Adult datasets. The results of varying τ are shown in Fig. 12. The values of ϵ was set to 7. Similar to the experimental results on the numerical dataset, the MSE of the proposed method is the smallest for any parameter setting. For the uniform dataset, each measured value differs from the true value, however, the true distribution remains uniform, implying that the frequency does not change much. Hence, the MSEs of the proposed method are similar to those of IDUE(O) (Fig. 12a). Results have shown that IDUE(O) achieves higher accuracy than IDUE(R) [18]. However, the effect of the measured value being different from the true value is more significant in the peak dataset. This is because the frequency of each value can be different. Therefore, the MSEs of the proposed method are less than those of other methods (Fig. 12b.) The degree of variation in the frequency of each value in the normal dataset is between the degrees of variation in the peak and uniform datasets, respectively. Hence, the difference between the MSEs of the proposed method and those of other methods in the normal dataset is larger and smaller than that in the uniform and peak datasets, respectively (Fig. 12c). For categorical attributes, the distribution of the values of AReM dataset is gentler than that of the values of Adult dataset. This characteristic is reflected in the results of their MSEs, as shown in (Figs. 12d and 12e).

The experimental results of uRAP and IDUE(O) are comparable because uRAP and IDUE(O) depend on the same OUE mechanism. The main task of PM-OPT is for federated learning; therefore, the accuracy of estimating the distribution of user data is not very high.

The results of varying ϵ are shown in Fig. 13. In the peak dataset, MSEs of the proposed method are similar to those of other methods when ϵ is large (Fig. 13a.) However, MSEs of the proposed method are the smallest for all other datasets (Figs. 13b–13e.) Similar to the results in Fig. 12, the greater the variations in the dataset, the more pronounced the effectiveness of the proposed method. This is because the effect of the measured error is much larger if the variations in the dataset are large.

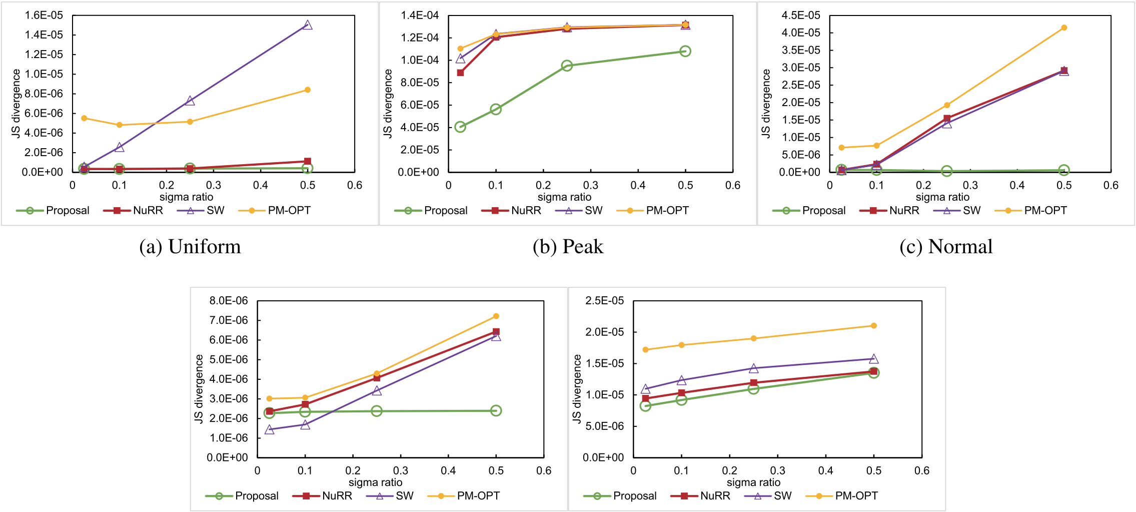
MSE is ideal for evaluating errors in large histogram values since it is highly sensitive to large histogram values. Therefore, MSE is suitable for the scenario where the analyzer wants to know the rough distribution of the data. However,



(a) Uniform (b) Peak (c) Normal

(d) ADL (e) Adult

FIGURE 13. MSEs of generated histograms on categorical datasets with varying ϵ.



(a) Uniform (b) Peak (c) Normal

(d) AReM (e) Adult

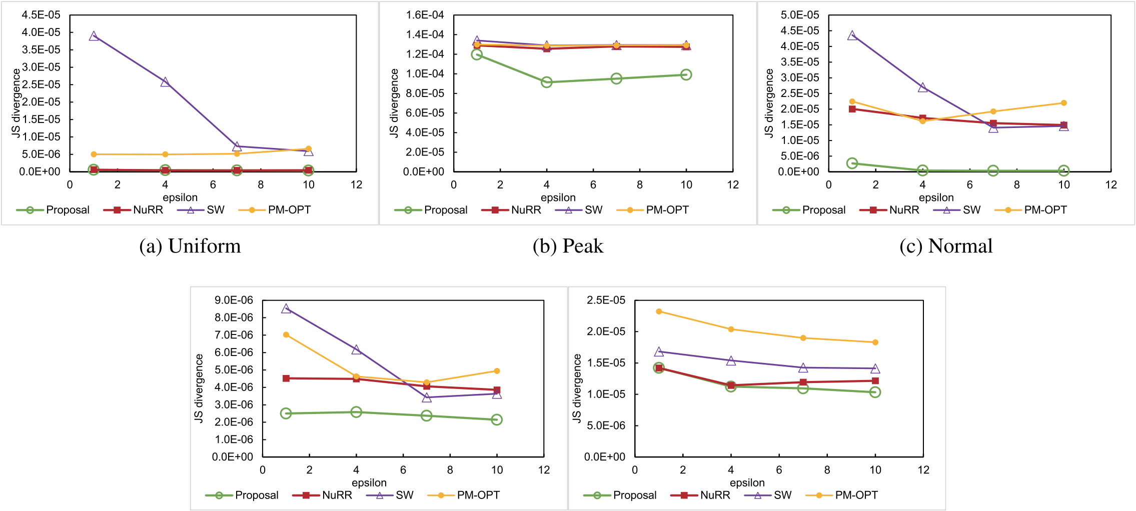
FIGURE 14. JS divergence of generated histograms on numerical datasets with varying the sigma ratio.

there is another scenario where we also want to know the finer details of the data distribution. In this case, Jensen–Shannon (JS) divergence [43] is suitable as a utility metric because it can evaluate the errors in small values of a histogram [44]. The results are shown in Figs. 14-17. The JS divergence results follow the same pattern as the MSE results. The JS divergence data, like the MSE results, reveal that our suggested method outperforms existing methods in terms of accuracy; however, the advantage of the proposed method has decreased marginally. In particular, the resulting histogram's error can be decreased by 40.4% on average in MSE and 29.6% on average in JS divergence. Therefore, we can infer that the proposed method is effective for both cases where a data analyst wants to know the underlying distribution of user data broadly, and the data analyst wants to know it in a fine-grained way.

To summarize the findings, the proposed method outperforms IDUE(R), IDUE(O), NuRR, uRAP, SW, and PM-OPT in terms of accuracy. This tendency was noticeable, particularly when there was a lot of sensor noise. The proposed technique considers sensing noise into account, and Algorithms 1 and 2 work efficiently limit the amount of noise imparted while maintaining the level of privacy protection.

G. CALCULATION COST

Local devices are used to run the anonymizing algorithms mentioned in Algorithms 1 and 2. We tested them on a laptop



(a) Uniform

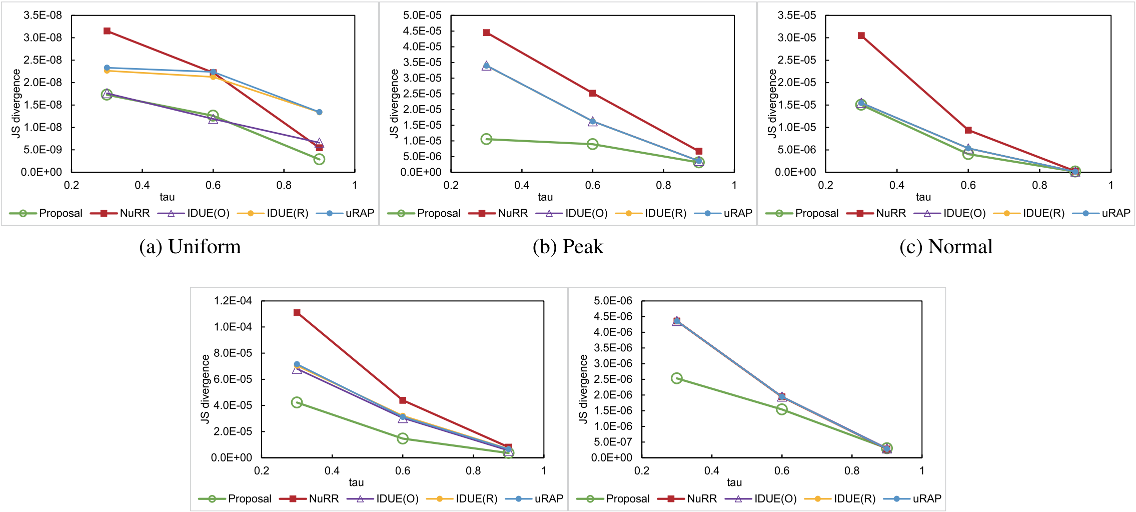
(b) Peak

(c) Normal

(d) AReM

(e) Adult

FIGURE 15. JS divergence of generated histograms on numerical datasets with varying ϵ.



(a) Uniform

(b) Peak

(c) Normal

(d) ADL

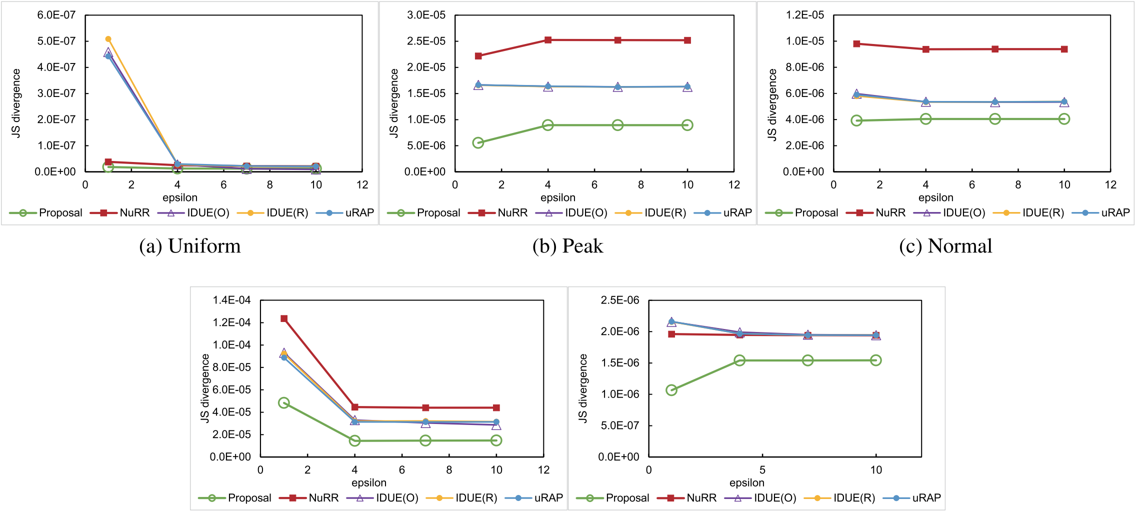
(e) Adult

FIGURE 16. JS divergence of generated histograms on categorical datasets with varying τ.

computer with 8GB of memory and a Core i5 10210U CPU. The method was executed in less than a second for each data set. A server is used to estimate the distribution of user data. We tested it in a workstation with 128GB memory and an Intel Xeon W-2295 CPU. It took 7.8 seconds on average for the AReM dataset with 42,239 values. The computation time of other methods is shorter because their algorithms are simple. Since the calculation time of our methods is proportional to the number of users and the number of attributes, it will take more time if the number of users increases. However, since collecting user data takes a certain amount of time, even if it takes a few minutes to estimate the data distribution, it is still considered practical enough.

V. RELATED RESEARCH WORK

A large number of research studies for anonymized data collection have been carried out. Wang *et al.* [45] proposed a method to identify the top-k most frequent new terms by collecting term usage data from each person under differential privacy. Kim *et al.* [46] derived population statistics by collecting differential private indoor positioning data. Anonymized data collection could also be realized based on encryption approaches [47], [48]. These methods can obtain aggregation values, and they do not aim to obtain each person's value. Moreover, these methods do not consider errors in values. On the other hand, in the proposed scenario, the aim is to obtain each person's value with as high accuracy as



(a) Uniform

(b) Peak

(c) Normal

(d) ADL

(e) Adult

FIGURE 17. **JS divergence of generated histograms on categorical datasets with varying ϵ.**

possible since services such as recommender systems need individuals' attribute values.

Abul *et al.* [49] and Sei and Ohsuga [50] proposed location anonymization methods taking into account location errors. These methods achieve k-anonymity, which is a basic privacy metric. However, they cannot be applied to ϵ-differential privacy.

Ge *et al.* [51], and Krishnan *et al.* [52] proposed methods to clean "dirty data" privately. They used differential privacy as a privacy metric and focused on data cleaning for resolving inconsistent attributes of an extensive database containing several people's true data. They assumed that each database value was true and used the Laplace mechanism without considering errors in values.

Several studies proposed machine learning methods, such as deep neural networks (deep learning) for IoT sensing values with differential privacy. Shi *et al.* [53] proposed a reinforcement technique for transportation network companies using passengers' data. Xu *et al.* [54] focused on mobile data analysis in edge computing. Guan *et al.* [55] applied machine learning for the Internet of medical things. Although they use differential privacy as a privacy metric, they do not consider the proposed TDP. By applying TDP, it is believed that the accuracy of their methods increases while maintaining their privacy protection levels.

VI. DISCUSSION

A. ERROR PROBABILITY DISTRIBUTION

Because our proposal for numerical attributes assumes that the error follows a normal distribution, the anonymizer must check whether the error follows a normal distribution and obtain the normal distribution's standard deviation. Notably, the proposed concept can be applied to the normal distribution and other distributions.

Many studies are based on the assumption that GPS location measurement errors follow a normal distribution, simulating these errors by generating noise that follows a bivariate normal distribution [56]–[58]. In addition to GPS measurements, many studies assume that measurement errors by sensors follow a normal distribution, and many studies have confirmed that measurement errors follow a normal distribution using actual data. Ferreira *et al.* measured various data from smart meters and generated error values for voltage and reactive power using a normal distribution in their experiments [59]. Sun *et al.* proposed a method to infer user intentions using spatio-temporal information and user behavior [60]. In their experiments, the measurement noises were drawn from a normal distribution. Xiao *et al.* proposed an RFID-based localization and tracking system, measuring and analyzing the error of the radar antenna, then describing the measurement error as following a normal distribution [61].

Many researchers such as [42], [62]–[64] also assume that sensing errors in IoT devices follow a normal distribution. Moreover, several researchers confirmed that actual sensing data follows a normal distribution. For example, Devon *et al.* collected 29,000 pieces of GPS data and illustrated that a normal distribution fits the data [65]. Wang *et al.* [66] observed that the pose tracking accuracy of the Microsoft Kinect 2, which can perform real-time gesture recognition, fits a normal distribution. Gao *et al.* [67] generated sensing samples based on a normal distribution of their experiments. Nguyen *et al.* [68] discussed how the measurement errors of sensing locations could affect mobile, robotic, and wireless sensor networks. In their proposed algorithm, the location error is modeled to follow a normal distribution. Using a real dataset with sensing errors, they showed that their algorithm achieves high performance.

Therefore, we can assume that the error probability fits a normal distribution in many cases. Usually, sensor vendors show a data sheet for each sensor product that contains information about the sensor's accuracy. There are several ways to express accuracy. Sometimes it is expressed using a standard deviation; in this case, the anonymizer uses the standard deviation of the normal distribution. If the accuracy is expressed using the average error (let m be the average), we can obtain the standard deviation of the normal distribution using the following equation:

$$\text{standard deviation} = m\sqrt{2\pi}. \tag{42}$$

In IoT systems, a machine learning technique that includes deep neural networks has also been used. Estimated values from deep neural networks might include estimation errors, and several researchers such as [69]–[71] reported that the estimation errors followed a normal distribution. If an anonymizer can obtain several training samples, the anonymizer can produce the error probability distribution and calculate its standard deviation. The Anderson-Darling test [72], which tests whether samples come from a normal distribution, can be used to check whether the error probability distribution follows a normal distribution.

Android OS provides APIs for location, speed, and bearing. In addition, the APIs for the data return the measurement values and their accuracy at 68% confidence.[3] In a normal distribution, 68% of the data falls within one standard deviation from the mean.

Errors always exist in the measurements regardless of how carefully and scientifically the measurements are performed. However, error analysis allows scientists to evaluate the degree of uncertainty. It has been proven that if the measurements have many small random error sources and negligible systematic errors, the measurements are normally distributed [16].

Although not all measurement errors follow a normal distribution, as mentioned above, many measurement errors are considered to follow a normal distribution. The method proposed here targets the case where measurement errors are considered to follow a normal distribution. On the other hand, our proposed novel concept, TDP, can be applied to any other error models. We hope that this paper is the first step toward error-aware differential privacy.

B. MACHINE LEARNING WITH NOISY IoT DATA

Several studies generated machine learning models from noisy IoT data, and the models could achieve high accuracy [73], [74]. In contrast, studies on differential privacy showed that if the value of the privacy budget is small, i.e., if the noise added using differential privacy techniques is large, the accuracy of machine learning models will deteriorate significantly [53]–[55].

In this paper, we have shown that by considering the observation noise of IoT data, the amount of noise to protect privacy can be reduced. Our proposed concept can be used with existing differentially private techniques. Therefore, the accuracy of the machine learning model can be improved while maintaining the same privacy protection level.

C. FUTURE WORK

Some mechanisms can achieve differential privacy other than the Laplace mechanism and the randomized response mechanism. Geng and Viswanath [75] proposed the staircase mechanism for numerical values and proved that it was better when ϵ was very large. Andrés *et al.* [76] proposed the geo-indistinguishable mechanism, especially for location information. The authors of this paper believe that TDP can be applied to other privacy metrics that manage probability distributions of attribute values. These privacy metrics include pk-anonymity [77], [78], which is a probability extension of the k-anonymity [79] model, and t-closeness [80], which is a refinement of k-anonymity.

Future work includes applying TDP to other mechanisms to achieve differential privacy and other privacy metrics.

In this paper, the normal distribution as the error distribution is considered. However, other error distributions can also be considered. By replacing $N(x; \sigma^2)$ with other error distributions, new algorithms for error distributions can be introduced.

In this paper, the target scenario proposed is the collection of anonymized data from each person individually. Privacy-preserving data publishing, which states that a data holder has much personal data and anonymizes and publishes them, is another important scenario in the research area of privacy; ϵ-differential privacy can be applied in this case. Furthermore, TDP can also be applied to ϵ-differential privacy in this case. A concrete discussion on these concepts is presented in future work.

VII. CONCLUSION

Differential privacy can protect user privacy by adding noise to a target value, which must be protected. Sensing values in IoT environments involve some errors; however, existing solutions have not taken sensing noise into account. In other words, present systems attempt to protect the detected value in the presence of sensing noise. On the contrary, our research aims at protecting the true value. Our technique modifies the amount of added noise based on the sensor noise model, whereas existing systems do not. Our strategy can lower the amount of noise introduced by the differential privacy technique by roughly 20%. As a result, the resulting histogram's mean square error and JS divergence can be lowered by 40.4% and 29.6% on average, respectively. A novel privacy metamodel called TDP is introduced and applied to differential privacy since the data owner or anonymizer might not know the true value. We validate this result on synthetic and five real data sets. This is the first research work that proposes and applies TDP. The authors expect many studies based on TDP in the near future.

REFERENCES

[1] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, and S. Andreescu, "Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges," in *Proc. IEEE Int. Conf. Services Comput.*, Jun. 2015, pp. 285–292.

[3] https://developer.android.com/reference/android/location/

[2] M. Munoz-Organero, G. A. Ramíez-González, P. J. Munoz-Merino, and C. D. Kloos, "A collaborative recommender system based on space-time similarities," *IEEE Pervasive Comput.*, vol. 9, no. 3, pp. 81–87, Jul. 2010.

[3] J. Torres-Sospedra, J. Avariento, D. Rambla, R. Montoliu, S. Casteleyn, M. Benedito-Bordonau, M. Gould, and J. Huerta, "Enhancing integrated indoor/outdoor mobility in a smart campus," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 11, pp. 1955–1968, Nov. 2015.

[4] J. Tang, S. Fu, X. Liu, Y. Luo, and M. Xu, "Achieving privacy-preserving and lightweight truth discovery in mobile crowdsensing," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 29, 2021, doi: 10.1109/TKDE.2021.3054409.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. (TCC)*, 2006, pp. 265–284.

[6] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, Aug. 2014.

[7] F. Liu, "Generalized Gaussian mechanism for differential privacy," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 747–756, Apr. 2019.

[8] X. Ren, C. M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and P. S. Yu, "Lopub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.

[9] N. Phan, X. Wu, H. Hu, and D. Dou, "Adaptive Laplace mechanism: Differential privacy preservation in deep learning," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 385–394.

[10] X. Zhang, R. Chen, J. Xu, X. Meng, and Y. Xie, "Towards accurate histogram publication under differential privacy," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 587–595.

[11] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," in *Proc. IEEE 26th Int. Conf. Data Eng. (ICDE)*, 2010, pp. 225–236.

[12] X. Ding, S. Sheng, H. Zhou, X. Zhang, Z. Bao, P. Zhou, and H. Jin, "Differentially private triangle counting in large graphs," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 19, 2021.

[13] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[14] Z. Huang and W. Du, "OptRR: Optimizing randomized response schemes for privacy-preserving data mining," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 705–714.

[15] G. B. Airy, *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Whitefish, MT, USA: Kessinger Publishing, 2007.

[16] J. R. Taylor, *Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd ed. Sausalito, CA, USA: Univ. Science Books, 1997.

[17] T. Murakami and Y. Kawamoto, "Utility-optimized local differential privacy mechanisms for distribution estimation," in *Proc. USENIX Secur. Symp.*, 2019, pp. 1877–1894.

[18] X. Gu, M. Li, L. Xiong, and Y. Cao, "Providing input-discriminative protection for local differential privacy," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 505–516.

[19] Z. Li, T. Wang, M. Lopuhaä-Zwakenberg, N. Li, and B. Škoric, "Estimating numerical distributions under local differential privacy," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2020, pp. 621–635.

[20] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam, "Local differential privacy-based federated learning for Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, Jun. 2021.

[21] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, and L. Shu, "Authentication protocols for Internet of Things: A comprehensive survey," *Secur. Commun. Netw.*, vol. 2017, pp. 1–41, Nov. 2017.

[22] L. Babun, K. Denney, Z. B. Celik, P. McDaniel, and A. S. Uluagac, "A survey on IoT platforms: Communication, security, and privacy perspectives," *Comput. Netw.*, vol. 192, Jun. 2021, Art. no. 108040.

[23] M. A. Husnoo, A. Anwar, R. K. Chakrabortty, R. Doss, and M. J. Ryan, "Differential privacy for IoT-enabled critical infrastructure: A comprehensive survey," *IEEE Access*, vol. 9, pp. 153276–153304, 2021.

[24] C. Ma, L. Yuan, L. Han, M. Ding, R. Bhaskar, and J. Li, "Data level privacy preserving: A stochastic perturbation approach based on differential privacy," *IEEE Trans. Knowl. Data Eng.*, early access, Dec. 21, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9658125/, doi: 10.1109/TKDE.2021.3137047.

[25] J. A. Onesimu, J. Karthikeyan, and Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services," *Peer-Peer Netw. Appl.*, vol. 14, no. 3, pp. 1629–1649, May 2021. [Online]. Available: https://link.springer.com/article/10.1007/s12083-021-01077-7

[26] F. M. Zahid and C. Heumann, "Multiple imputation with sequential penalized regression," *Stat. Methods Med. Res.*, vol. 28, no. 5, pp. 1311–1327, May 2019.

[27] A. Lyon, "Why are normal distributions normal?" *Brit. J. Philosophy Sci.*, vol. 65, no. 3, pp. 621–649, Sep. 2014.

[28] 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.

[29] S. Winitzki, "A handy approximation for the error function and its inverse," Univ. Heidelberg, Heidelberg, Germany, Tech. Rep., 2008.

[30] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on MacOS 10.12," 2017, pp. 1–12, *arXiv:1709.02753*.

[31] Differential Privacy Team Apple, "Learning with privacy at scale," *Mach. Learn. J.*, vol. 1, no. 8, pp. 1–25, 2017.

[32] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM CCS*, 2014, pp. 1054–1067.

[33] D. Dua and C. Graff. (2019). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[34] F. Palumbo, C. Gallicchio, R. Pucci, and A. Micheli, "Human activity recognition using multisensor data fusion based on reservoir computing," *J. Ambient Intell. Smart Environ.*, vol. 8, no. 2, pp. 87–107, 2016.

[35] F. J. Ordóñez, P. de Toledo, A. Sanchis, F. J. Ordóñez, P. De Toledo, and A. Sanchis, "Activity recognition using hybrid generative/discriminative models on home environments using binary sensors," *Sensors*, vol. 13, no. 5, pp. 5460–5477, 2013.

[36] R. L. S. Torres, D. C. Ranasinghe, Q. Shi, and A. P. Sample, "Sensor enabled wearable RFID technology for mitigating the risk of falls near beds," in *Proc. IEEE Int. Conf. RFID (RFID)*, Apr. 2013, pp. 191–198.

[37] B. Kaluža, V. Mirchevska, E. Dovgan, M. Luštrek, and M. Gams, "An agent-based approach to care in independent living," in *Proc. Int. Joint Conf. Ambient Intell.*, 2010, pp. 177–186.

[38] X. Wang, V. Ly, G. Lu, and C. Kambhamettu, "Can we minimize the influence due to gender and race in age estimation?" in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, Dec. 2013, pp. 309–314.

[39] S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim, "Age estimation using a hierarchical classifier based on global and local facial features," *Pattern Recognit.*, vol. 44, no. 6, pp. 1262–1281, 2011.

[40] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[41] T. Wang, J. Blocki, N. Li, T. Wang, J. Blocki, and N. Li, "Locally differentially private protocols for frequency estimation," in *Proc. USENIX Secur. Symp.*, 2017, pp. 729–745.

[42] Y. Sei and A. Ohsuga, "Differentially private mobile crowd sensing considering sensing errors," *Sensors*, vol. 20, no. 10, pp. 2785:1–2785:25, May 2020.

[43] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[44] T. Murakami, H. Hino, and J. Sakuma, "Toward distribution estimation under local differential privacy with small samples," *Proc. Privacy Enhancing Technol.*, vol. 2018, no. 3, pp. 84–104, Jun. 2018.

[45] N. Wang, X. Xiao, Y. Yang, T. D. Hoang, H. Shin, J. Shin, and G. Yu, "PrivTrie: Effective frequent term discovery under local differential privacy," in *Proc. IEEE 34th Int. Conf. Data Eng. (ICDE)*, Apr. 2018, pp. 821–832.

[46] J. W. Kim, D.-H. Kim, and B. Jang, "Application of local differential privacy to collection of indoor positioning data," *IEEE Access*, vol. 6, pp. 4276–4286, 2018.

[47] S. Tonyali, K. Akkaya, N. Saputro, A. S. Uluagac, and M. Nojoumian, "Privacy-preserving protocols for secure and reliable data aggregation in IoT-enabled smart metering systems," *Future Gener. Comput. Syst.*, vol. 78, pp. 547–557, Jan. 2018.

[48] Y. Liu, W. Guo, C.-I. Fan, L. Chang, and C. Cheng, "A practical privacy-preserving data aggregation (3PDA) scheme for smart grid," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1767–1774, Mar. 2018.

[49] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 376–385.

[50] Y. Sei and A. Ohsuga, "Location anonymization with considering errors and existence probability," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 12, pp. 3207–3218, Dec. 2017.

[51] C. Ge, I. F. Ilyas, X. He, and A. Machanavajjhala, "Private exploration primitives for data cleaning," 2017, pp. 1–17, *arXiv:1712.10266*.

[52] S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, and T. Kraska, "Privateclean: Data cleaning and differential privacy," in *Proc. ACM SIGMOD*, 2016, pp. 937–951.

[53] D. Shi, J. Ding, S. M. Errapotu, H. Yue, W. Xu, X. Zhou, and M. Pan, "Deep *Q*-network-based route scheduling for TNC vehicles with Passengers' location differential privacy," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7681–7692, Oct. 2019.

[54] C. Xu, J. Ren, L. She, Y. Zhang, Z. Qin, and K. Ren, "EdgeSanitizer: Locally differentially private deep inference at the edge for mobile data analytics," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5140–5151, Jun. 2019.

[55] Z. Guan, Z. Lv, X. Du, L. Wu, and M. Guizani, "Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach," *Future Gener. Comput. Syst.*, vol. 98, pp. 60–68, Sep. 2019.

[56] P. Chao, W. Hua, R. Mao, J. Xu, and X. Zhou, "A survey and quantitative study on map inference algorithms from GPS trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 15–28, Jan. 2022.

[57] E. Frentzos, K. Gratsias, and Y. Theodoridis, "On the effect of location uncertainty in spatial querying," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 3, pp. 366–383, Mar. 2009.

[58] D. Zhang, Z. Chang, S. Wu, Y. Yuan, K.-L. Tan, and G. Chen, "Continuous trajectory similarity search for online outlier detection," *IEEE Trans. Knowl. Data Eng.*, early access, Dec. 24, 2020, doi: 10.1109/TKDE.2020.3046670.

[59] T. S. D. Ferreira, F. C. L. Trindade, and J. C. M. Vieira, "Load flow-based method for nontechnical electrical loss detection and location in distribution systems using smart meters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 3671–3681, Sep. 2020.

[60] Y. Sun, N. J. Yuan, X. Xie, K. McDonald, and R. Zhang, "Collaborative intent prediction with real-time contextual data," *ACM Trans. Inf. Syst.*, vol. 35, no. 4, pp. 1–33, Aug. 2017.

[61] F. Xiao, Z. Wang, N. Ye, R. Wang, and X.-Y. Li, "One more tag enables fine-grained RFID localization and tracking," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 161–174, Feb. 2018.

[62] R. Peng and M. L. Sichitiu, "Angle of arrival localization for wireless sensor networks," in *Proc. 3rd Annu. IEEE Commun. Soc. Sensor Ad Hoc Commun. Netw. (SECON)*, 2006, pp. 374–382.

[63] I. Floris, P. A. Calderón, S. Sales, and J. M. Adam, "Effects of core position uncertainty on optical shape sensor accuracy," *Measurement*, vol. 139, pp. 21–33, Jun. 2019.

[64] A. Burguera, Y. González, and G. Oliver, "Sonar sensor models and their application to mobile robot localization," *Sensors*, vol. 9, no. 12, pp. 10217–10243, Dec. 2009.

[65] D. DeVon, T. Holzer, and S. Sarkani, "Minimizing uncertainty and improving accuracy when fusing multiple stationary GPS receivers," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2015, pp. 83–88.

[66] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Evaluation of pose tracking accuracy in the first and second generations of Microsoft Kinect," in *Proc. Int. Conf. Healthcare Informat. (ICHI)*, Oct. 2015, pp. 380–389.

[67] X. Gao, S. Chen, and G. Chen, "MAB-based reinforced worker selection framework for budgeted spatial crowdsensing," *IEEE Trans. Knowl. Data Eng.*, early access, May 4, 2020, doi: 10.1109/TKDE.2020.2992531.

[68] L. V. Nguyen, S. Kodagoda, R. Ranasinghe, and G. Dissanayake, "Adaptive placement for mobile sensors in spatial prediction under locational errors," *IEEE Sensors J.*, vol. 17, no. 3, pp. 794–802, Feb. 2017.

[69] P. N. P. Barbeiro, J. Krstulovic, H. Teixeira, J. Pereira, F. J. Soares, and J. P. Iria, "State estimation in distribution smart grids using autoencoders," in *Proc. IEEE 8th Int. Power Eng. Optim. Conf. (PEOCO)*, Mar. 2014, pp. 358–363.

[70] A. Magaña, H. Wu, P. Bauer, and G. Reinhart, "PoseNetwork: Pipeline for the automated generation of synthetic training data and CNN for object detection, segmentation, and orientation estimation," in *Proc. 25th IEEE Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, Sep. 2020, pp. 587–594.

[71] M. Uss, B. Vozel, V. Lukin, and K. Chehdi, "Efficient discrimination and localization of multimodal remote sensing images using CNN-based prediction of localization uncertainty," *Remote Sens.*, vol. 12, no. 4, p. 703, Feb. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/4/703

[72] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons," *J. Amer. Stat. Assoc.*, vol. 69, no. 347, pp. 730–737, Sep. 1974.

[73] M. Li, L. Xie, Z. Lv, J. Li, and Z. Wang, "Multistep deep system for multimodal emotion detection with invalid data in the Internet of Things," *IEEE Access*, vol. 8, pp. 187208–187221, 2020.

[74] C. Lee, J. Lin, P. Chen, and Y. Chang, "Deep learning-constructed joint transmission-recognition for Internet of Things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.

[75] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 925–951, Feb. 2016.

[76] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. ACM CCS*, 2013, pp. 901–914.

[77] E. Kimura, K. Chida, D. Ikarashi, K. Hamada, and K. Ishihara, "Statistical disclosure limitation of health data based on Pk-anonymity," *Stud. Health Technol. Informat.*, vol. 180, pp. 1117–1119, Aug. 2012.

[78] M. Kakizawa, C. Watanabe, R. Furukawa, and T. Takahashi, "Improvement of Pk-anonymization," in *Proc. IEEE 33rd Int. Symp. Reliable Distrib. Syst. Workshops*, Oct. 2014, pp. 82–87.

[79] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD*, 2005, pp. 49–60.

[80] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

YUICHI SEI (Member, IEEE) received the Ph.D. degree in information science and technology from The University of Tokyo, in 2009. From 2009 to 2012, he was with Mitsubishi Research Institute. He joined The University of Electro-Communications, in 2013, and is currently an Associate Professor with the Graduate School of Informatics and Engineering. He is also a Visiting Researcher at Mitsubishi Research Institute and an Adjunct Researcher at Waseda University. His current research interests include pervasive computing, privacy-preserving data mining, and software engineering. He was a recipient of the IPSJ Best Paper Award and the JSCE Hydraulic Engineering Best Paper Award, in 2017.

AKIHIKO OHSUGA (Member, IEEE) received the Ph.D. degree in computer science from Waseda University, in 1995. From 1981 to 2007, he was with Toshiba Corporation. He joined The University of Electro-Communications, in 2007, and is currently a Professor with the Graduate School of Informatics and Engineering. He is also the Dean of the Graduate School of Information Systems. He is also a Visiting Professor at the National Institute of Informatics. His research interests include agent technologies, web intelligence, and software engineering. He is a member of IEEE Computer Society (IEEE CS), Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society for Artificial Intelligence (JSAI), Japan Society for Software Science and Technology (JSSST), and the Institute of Electrical Engineers of Japan (IEEJ). He received the IPSJ Best Paper Awards in 1987 and 2017. He was the Chair of IEEE CS Japan Chapter. He was a member of the Board of Directors of *JSAI* and *JSSST*.