IEEE Open Journal of the Computer Society

Received 31 October 2022; accepted 10 December 2022. Date of publication 14 December 2022; date of current version 23 December 2022. The review of this article was arranged by Associate Editor Jiankun Hu.

Digital Object Identifier 10.1109/OJCS.2022.3228999

Local Differential Privacy for Person-to-Person Interactions

YUICHI SEI [1, 2] (Member, IEEE), AND AKIHIKO OHSUGA [1] (Member, IEEE)

[1]The University of Electro-Communications, Tokyo 182-8585, Japan
[2]JST, PRESTO, Kawaguchi 332-0012, Saitama, Japan

CORRESPONDING AUTHOR: YUICHI SEI (e-mail: seiuny@uec.ac.jp)

This work was supported in part by the JSPS KAKENHI under Grants JP21H03496 and JP22K12157, and in part by the JST, PRESTO under Grant JPMJPR1934.

ABSTRACT Currently, many global organizations collect personal data for marketing, recommendation system improvement, and other purposes. Some organizations collect personal data securely based on a technique known as ϵ-local differential privacy (LDP). Under LDP, a privacy budget is allocated to each user in advance. Each time the user's data are collected, the user's privacy budget is consumed, and their privacy is protected by ensuring that the remaining privacy budget is greater than or equal to zero. Existing research and organizations assume that each individual's data are completely unrelated to other individuals' data. However, this assumption does not hold in a situation where interaction data between users are collected from them. In this case, each user's privacy is not sufficiently protected because the privacy budget is actually overspent. In this study, the issue of local differential privacy for person-to-person interactions is clarified. We propose a mechanism that satisfies LDP in a person-to-person interaction scenario. Mathematical analysis and experimental results show that the proposed mechanism can maintain high data utility while ensuring LDP compared to existing methods.

INDEX TERMS Ethics and privacy, local differential privacy, person-to-person interaction.

I. INTRODUCTION

Although the analysis of Big Data, including personal data, has enabled the emergence of new services, privacy information leakage is a serious issue [1], [2]. Several large organizations such as Apple, Google, and Microsoft have been collecting users' information while protecting their privacy [3], [4], [5], [6] using the ϵ-local differential privacy (LDP) technique [7]. Although LDP is considered the best technology for privacy protection [8], [9], these organizations additionally apply explicit privacy policies for data collection. For example, Apple collects data from users regarding their emoji usage through LDP; however, it does not collect the users' identities.

In LDP, each user is assigned a privacy budget, which is a non-negative real value. When the user data are sent to the data collector, an amount (or the entirety) of the privacy budget of the user is consumed. The total privacy budget and the value of each privacy budget consumed can be controlled by an agreement between the data collector and user. For example, suppose that the privacy budget for user A is 10.0, and each privacy budget value consumed to transmit the data of this user is 1.0. The data collector can retrieve user A's data 10 times. To ensure continuous data collection, the total privacy budget of each user is regularly restored.

If the data collected by the data collector refer to a user's information regardless of other users, there are no issues because the user has already agreed to the privacy policy. However, what happens if the data collected refer to person-to-person interaction information? Suppose that user A sends an email to user B, and the data collector gathers word usage information through LDP under the privacy policy agreed upon with user A. The data collected are the word information used by user A, but for user B, the data are the word information they have received. In other words, it is equivalent to collecting user B's data. Therefore, the data collector must also consider user B's privacy. However, currently, it is not checked whether user B has agreed to the privacy policy or not. Even if user B has agreed to this policy, no one has control over user's B privacy budget.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

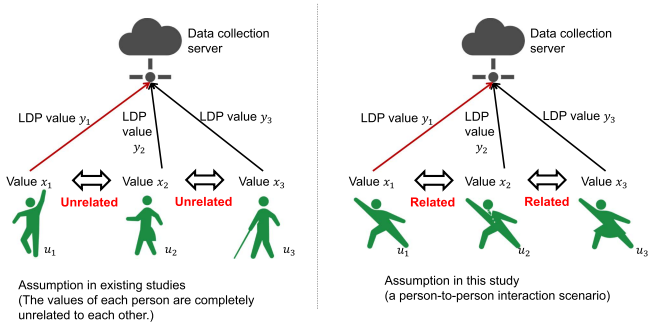


FIGURE 1. Assumptions in existing studies and this study.

Fig. 1 shows the difference between the assumptions in existing studies and this study. When user u_1 sends LDP value y_1 of their true value x_1 to the data collection server, only u_1's private information is provided to the data collection server in existing studies. This is because the values of each user are completely unrelated to each other. Moreover, suppose that each value depends on the other. In this case, when user u_1 sends LDP value y_1, u_1, u_2, and u_3's information are provided to the data collection server. In other words, although u_2 does not send any information to the data collection server, through the behavior of u_1, partial information of u_2 is provided to the data collection server.

In this study, this problem was formalized as a person-to-person interaction in LDP. To concentrate the discussion in this paper on the new concept of person-to-person interaction under ϵ-LDP, we targeted the relatively simple task of obtaining average values from users. The recommendations in this paper are expected to have a considerable impact on all organizations that collect person-to-person interaction data using ϵ-LDP.

The main contributions of this paper are as follows. First, we formalized a person-to-person interaction problem under ϵ-LDP, and we demonstrated that several existing studies on LDP do not ensure ϵ-LDP in a person-to-person interaction scenario. Second, we proposed a method that can ensure ϵ-LDP. Third, the approximate amount of error between the true average value and the average value estimated by the proposed method was analyzed mathematically. Fourth, based on the results of extensive simulations, we showed that the proposed method outperformed existing methods and that the results of the mathematical analysis were in perfect agreement with the experimental results.

II. RELATED WORK AND REAL APPLICATIONS

A. LOCAL DIFFERENTIAL PRIVACY

In technical terms, LDP is defined as ϵ-LDP, where parameter ϵ represents a privacy budget. There are several relaxation concepts of ϵ-LDP such as (ϵ, δ)-LDP and Renyi differential privacy [10]. Although the concepts discussed in this paper can be applied to the relaxations of LDP, we focused on ϵ-LDP to simplify the discussion. ϵ-LDP is defined as follows.

Definition II.1 (ϵ-LDP): Let X represent a domain of the data of a user and let Y be an arbitrary set. A randomized mechanism M provides ϵ-LDP if and only if for any $x, x' \in X$ and any $y \in Y$,

$$P(M(x) = y) \leq e^{\epsilon} P(M(x') = y). \tag{1}$$

Several techniques have been proposed to achieve LDP. One of the most commonly used techniques is the Laplace mechanism [11]. To introduce the Laplace mechanism, the concept of global sensitivity is defined.

Definition II.2 (Global sensitivity): For a function $f: X \to Y$, the global sensitivity of f is defined as follows.

$$\Delta f = \max_{x,x' \in X} |f(x) - f(x')|. \tag{2}$$

Theorem II.3 (Laplace mechanism [11]): Let Δf be the global sensitivity of a function $f : X \to Y$ and let $\mathcal{L}(v)$ represent the Laplace distribution, with the mean as zero and the scale parameter as v. The following mechanism M ensures ϵ-LDP.

$$M(x) = f(x) + \mathcal{L}\left(\frac{\Delta f}{\epsilon}\right). \tag{3}$$

Many methods have been proposed for estimating a histogram distribution of users' values under ϵ-LDP, such as the Randomized Aggregatable Privacy-Preserving Ordinal Response, *Sarve*, and so on [4], [12]. Although such methods achieve high accuracy, their techniques cannot be applied to a person-to-person interaction scenario. This is because they assume that each user's value is not dependent on another user's value.

There are also several methods for estimating the average value of users. Xue et al. proposed (τ, ϵ)-personalized local differential privacy (PLDP) as a privacy metric, Duchi's solution with PLDP (DCP), and piecewise mechanism with PLDP (PWP) [13]. The (τ, ϵ)-PLDP is a privacy metric that weakens ϵ-LDP, but DCP and PWP can be used for ϵ-LDP. We can assume that the range of a value is $[-1, 1]$ without loss of generality. In DCP, each user sends a randomized value v with a probability

$$Pr(\epsilon, x) = \frac{(e^{\epsilon} - 1) \cdot x}{(e^{\epsilon} + 1) \cdot 2} + \frac{1}{2}. \tag{4}$$

In PWP, each user randomly selects a value from a range around the true value with probability p, where the value of p is determined from ϵ. With probability $1 - p$, a value from a wider range is randomly selected, and the selected value is sent to the server. Because the ratio of $p/(1 - p)$ is e^{ϵ}, PWP ensures ϵ-LDP.

Li et al. proposed a square wave mechanism (SW) [14]. This mechanism is similar to PWP, but the range of LDP values to be selected is different.

Many other LDP methods have been proposed. Navidan et al. proposed a framework that estimates the number of people in each area while protecting each user's location privacy using local differential privacy [15]. In the framework, users measure the Received Signal Strength Indicator (RSSI) and

TABLE 1. Notations

$X_{i,j}$	Domain of interaction data between u_i and u_j
X_i	Domain of u_i's data
$x_{i,j}$	True value of the interaction between u_i and u_j
x_i	True value of u_i, obtained from $x_{i,j} \in X_{i,j}$ for all j (except for $i = j$).
$\Delta f_{i,j}$	Global sensitivity of $x_{i,j}$
Δf_i	Global sensitivity of x_i
$\epsilon_{i,j}$	Privacy budget for $x_{i,j}$
ϵ_i	Privacy budget for x_i
$\mathcal{L}(v)$	Function of Laplace distribution, with the mean as zero and the scale parameter as v

determine their locations based on the RSSI. The users then perturb their location information and send it to the data aggregator, who estimates how many users are in each location. The experimental results showed that the proposed framework could estimate location frequency while ensuring differential privacy.

Kim and Jang [16] proposed a data collection method of workload-aware differentially private positioning. They assume that location is hierarchical and aim to estimate the density at each location for each level of the hierarchy by utilizing local differential privacy. Their method can provide an optimal perturbation scheme to minimize the estimation error for a given workload.

Although many studies target one-shot data-sharing scenarios, several studies consider data stream cases. Please note that our proposed method can be used for data stream cases by dividing the privacy budget by the number of data acquisitions. By using methods for specified data stream cases, the accuracy of the data analysis can be enhanced. For example, Ren et al. [17] proposed an LDP mechanism for infinite data stream that targets w-event privacy, which ensures LDP for arbitrary time windows consisting of consecutive w time steps. In the future, we will propose a specialized method for measuring time series' data.

Ren et al. [18] proposed an anonymous data aggregation scheme that allows the server to estimate the number of users located within each value area without knowing the location of individual users. In particular, the authors focus on high-dimensional values. The domain sizes of datasets used in the experiments in [18] were 2^{16}, 2^{52}, and 2^{77}. Future experiments with such high-dimensional data sets should be conducted to test our proposed method.

These studies are excellent, but do not take user interaction into account.

In recent years, studies on federated learning with LDP have gained attention [19], [20], [21]. In typical federated learning using LDP, the server sends the machine-learning model parameters under training to the clients. Each client independently trains the machine-learning model using private local data samples. The updated gradient information is sent to the server under the protection of LDP. If each private local data sample is completely unrelated to the private local data samples of other users, ϵ-LDP can be ensured in these studies. However, for the person-to-person interaction data envisioned in this study, when one user sends information to a server through LDP, there is a need to consider the loss of privacy of other users as well.

All the extant studies on LDP such as [4], [9], [22], [23] assume that one user's value is independent of that of any other user. In many cases, this assumption is correct. However, in some scenarios, this assumption does not hold, as discussed in Section II-B.

Example 1: Alice transferred \$50 to Bob in a day. Alice has agreed to a 10-LDP (i.e., the amount of the privacy budget is 10), which allows a data collector to gather the amount per day transferred by Alice. Based on this policy, Alice sends the LDP value (e.g., \$53) to the data collector by consuming a 10-privacy budget. Since Alice's identity is not sent to the data collector, it only knows that someone transferred \$53 on that day.

In the above example, the information sent is related to Alice's money transfer. However, for Bob, the information sent is related to Bob's receipt of money. In this case, Alice's 10-privacy budget and Bob's 10-privacy budget are consumed. Therefore, if Bob's transmission information is also collected, the total amount of privacy budget consumed will be 20, surpassing the upper limit of 10. Such problems occur in person-to-person interactions in LDP.

B. APPLICATION OF LDP UNDER PERSON-TO-PERSON INTERACTIONS

Recently, LDP has been widely applied to many real services. Apple collects pictogram usage information from a user under LDP to analyze the use frequency of each pictogram [3]. However, Apple does not seem to care about the receiver's privacy.

Several email datasets contain anonymized text information and pseudo personal, sender, and receiver IDs [24]. Such data can be collected under LDP from each user. Emails are generally considered personal data that must be handled with care, regardless of the data that are sent or received. Therefore, if the email information of a sender is collected under LDP, this collection should consume the privacy budget of not only the sender but also the receiver.

Human relationship information such as information from online social networks is another form of privacy information. There are several anonymized datasets on human relationships, such as Epinions social network [25]. If the data collector gathers information about who a user is connected to and trusts, the privacy budget of not only the user but also the other person must be consumed.

III. PROBLEM DEFINITION

We defined the problem of LDP for person-to-person interactions. This scenario was not assumed in existing studies, but it is present in real-world scenarios. One of the most important contributions of this work is to clarify this problem. Numerous forms of person-to-person interactions are possible, but to

simplify the discussion in this paper, we will limit our analysis to the following interactions.

Definition III.1 (ϵ-LDP in a person-to-person interaction scenario): Let X_i represent a domain of user u_i's data and let $X_{i,j}$ represent a domain of interaction data between two users u_i and u_j $(i, j = 1, \ldots, n\ (i \neq j))$. The value of $x_i \in X_i$ is obtained from $x_{i,j} \in X_{i,j}$ for all j except for $i = j$; i.e., $x_i = f(x_{i,1}, \ldots, x_{i,i-1}, x_{i,i+1}, \ldots, x_{i,n})$ for a function $f : X_{i,j}^{n-1} \to X_i$.

User u_i sends information x_i under ϵ-LDP using mechanism M, which is defined in Definition II.1.

Theorem III.2 (Consumed privacy budget of ϵ-LDP for person-to-person interactions): In a scenario of ϵ-LDP for person-to-person interactions, the consumed privacy budget of user u_i is ϵ. The privacy budget of user u_j is also consumed, and this amount is represented by

$$\begin{aligned} &\min \epsilon_j, s.t.\ \ P(M(f(x_{i,1}, \ldots, x_{i,n})) = y) \\ &\leq e^{\epsilon_j} P\left(M(f(\ldots, x_{i,j-1}, x_{i,j}{}', x_{i,j+1}, \ldots)) = y\right), \end{aligned} \tag{5}$$

for any $x_{i,j}, x'_{i,j} \in X_{i,j}$.

Proof: Regarding user u_i, the consumed privacy budget is ϵ because x_i is collected under ϵ-LDP.

For user u_j $(j \neq i)$, the following expression should be satisfied for any $x_{i,j}, x'_{i,j} \in X_{i,j}$ to ensure ϵ_j-LDP because of Definition II.1.

$$\begin{aligned} &P(M(f(x_{i,1}, \ldots, x_{i,n})) = y) \\ &\leq e^{\epsilon_j} P(M(f(x_{i,1}, \ldots, x_{i,j-1}, x'_{i,j}, x_{i,j+1}, \ldots, x_{i,n})) = y). \end{aligned} \tag{6}$$

The smaller the value of ϵ_j, the smaller the privacy budget consumed and the more robustly the privacy is protected. Therefore, the consumed privacy budget is the minimum value that satisfies (6). ■

The problem definition in this paper is as follows.

Problem III.3 (Obtaining the average value under ϵ-LDP in a person-to-person interaction scenario): Assume there are n users $(u_1, \ldots, u_n)$, and each privacy budget is set to ϵ_i. In a person-to-person interaction scenario, the average value of $x_1, \ldots, x_n$ is obtained with high accuracy while ensuring ϵ_i-LDP for each user u_i.

Note that we do not propose a new privacy metric, but we strictly follow ϵ-LDP. The difference between the target of this paper and the existing studies is whether or not each user's data contain information on other users' data, which should be protected. To simplify the discussion, the target of this analysis is to obtain the average value of all users' data. However, the concept of ϵ-LDP in a person-to-person interaction scenario can be applied to any other analysis such as histogram estimation and machine learning. Such analysis remains to be undertaken in future work.

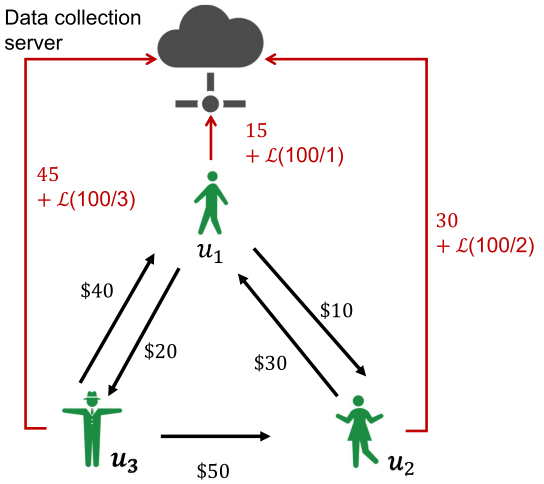


FIGURE 2. A directed graph of Examples 2 and 3.

IV. PROPOSED METHOD

The main notations used in this study are listed in Table 1. We mainly used a Laplace mechanism. To use this mechanism, the global sensitivity of each user should be clarified.

Definition IV.1 (Global sensitivity for a person-to-person interaction): For user u_i, the global sensitivity is the same as that given in Definition II.2.

For user $u_j (j \neq i)$, the global sensitivity of f is defined as

$$\Delta f_{i,j} = \max_{x_{i,j}, x'_{i,j} \in X_{i,j}} |f(\ldots, x_{i,j}, \ldots) - f(\ldots, x'_{i,j}, \ldots)|. \tag{7}$$

Theorem IV.2 (Consumed privacy budget of the Laplace mechanism in a person-to-person interaction): Suppose user u_i sends the value of x_i to the data collector under ϵ_i-LDP using a Laplace mechanism. Let Δf_i represent the global sensitivity of x_i and let $\Delta f_{i,j}$ represent the global sensitivity of $x_{i,j}$. In this case, ϵ_i of the privacy budget of user u_i and $\epsilon_i \Delta f_{i,j} / \Delta f_i$ of the privacy budget of everyone else $p_{i,j}$ are consumed.

Proof: For x_i, this mechanism ensures ϵ_i-LDP according to (1).

For $x_{i,j}$, the global sensitivity is $\Delta_{i,j}$. The value sent to the server is represented by

$$f(x_i) + \mathcal{L}\left(\frac{\Delta f_i}{\epsilon_i}\right) = f(x_i) + \mathcal{L}\left(\frac{\Delta f_{i,j}}{\epsilon_i \Delta f_{i,j} / \Delta f_i}\right). \tag{8}$$

Therefore, according to (1), this mechanism ensures that $(\epsilon_i \Delta f_{i,j} / \Delta f_i)$-LDP regarding $x_{i,j}$. ■

Example 2: Consider people u_1, u_2, and u_3 giving money to each other. The maximum amount of money given is limited to \$100. Therefore, $\Delta f_1 = \Delta f_2 = \Delta f_3 = 100$. User u_1 gave \$10 and \$20 to u_2 and u_3, respectively. User u_2 gave \$30 to u_1. User u_3 gave \$40 and \$50 to u_1 and u_2, respectively (see Fig. 2). User u_1 sends information about how many dollars u_1 gave on average to the server. In this case, $x_1 = f(x_{1,2}, x_{1,3}) = 15$ where function f is a function that

calculates the average. In this example, $\Delta f_{i,j} = 50$ because a change of $x_{i,j}$ can affect the value of x_i by up to 50.

When user u_1 sends a value under 1-LDP to the server, i.e., the result of 15+$\mathcal{L}$(100/1) is sent to the server, this behavior consumes 1, 0.5, and 0.5 of the privacy budget of users u_1, u_2, u_3, respectively.

So far, we assumed that only one user (u_i) sends their LDP value to the data collection server. When several users send their LDP values, the interaction-composition should be considered.

Theorem IV.3 (Interaction-composition property of LDP in person-to-person interactions): Suppose that the private information of $u_i (i = 1, \ldots, n)$ is collected under ϵ_i-LDP by the data collection server. Let $\epsilon_{i,j}$ represent the consumed privacy budget of user u_j by the collection of the private information of u_i. In this case, the total privacy budget consumed for u_i is represented by

$$\widehat{\epsilon_i} = \epsilon_i + \sum_{j \neq i} \epsilon_{j,i} \Delta f_{j,i} / \Delta f_j. \tag{9}$$

Proof: The information related to u_i is represented by

$$\begin{cases} x_{i,1}, \ldots, x_{i,i-1}, x_{i,i+1}, \ldots, x_{i,n} \\ x_{1,i}, \ldots, x_{i-1,i}, x_{i+1,i}, \ldots, x_{n,i}. \end{cases} \tag{10}$$

The value of x_i is calculated based on the upper part of (10); that is, $x_i = f(x_{i,1}, \ldots, x_{i,i-1}, x_{i,i+1}, \ldots, x_{i,n})$. This value is sent to the server under the privacy budget ϵ_i.

Each value $x_{j,i}$ of the lower part of (10) is sent by user u_j under the privacy budget $\epsilon_{j,i}$ for $x_{j,i}$. Because of the sequential composition property of differential privacy [26], the total privacy loss is calculated by (9). ■

Example 3: Consider the same case described in Example 2. The values of x_1, x_2, x_3 are 15, 30, and 45, respectively. Consider people u_1, u_2, and u_3 sending their values x_1, x_2, x_3 under 1-LDP, 2-LDP, and 3-LDP, respectively. In this case, in each report of user u_i, the total privacy loss of u_1, u_2, and u_3 are (1+2/2+3/2 = 3.5), (2+1/2+3/2 = 4), and (3+1/2+2/2 = 4.5), respectively.

So far, we discussed generalized scenarios where the global sensitivity and privacy budget are different for each user. Usually, however, these values are common for all users. In this case, the following theorem holds.

Theorem IV.4: Consider that there are n users and each user u_i sends x_i under ϵ-LDP. In this case, each transfer of data of u_i consumes $\epsilon/(n-1)$ of the privacy budget of another user. The total privacy loss of each user u_i is represented by $\epsilon + \sum_{j \neq i} \epsilon/(n-1) = 2\epsilon$.

In the following text, the expected amount of error of the estimated mean under ϵ-LDP in a person-to-person interaction is discussed. We assume that each privacy budget of each user is ϵ and the global sensitivity for each is Δf. Let $\mathcal{L}(x; s)$ represent the probability density function (PDF) of the Laplace distribution with mean 0 and scale s. The probability distribution of the sum of n Laplace random variables is represented by the following equation.

$$\begin{aligned} \mathcal{L}_n(x; s) &= \int_{x_1=-\infty}^{\infty} \cdots \int_{x_{n-1}=-\infty}^{\infty} \\ &\mathcal{L}(x_1; s) \cdots \mathcal{L}(x_{n-1}; s) \mathcal{L}\left(x - \sum_{i=1}^{n-1} x_i; s\right) dx_1 \cdots dx_{n-1} \\ &= \frac{e^{-\frac{|x|}{s}} \sum_{i=0}^{n-1} a_{n,i} s^i |x|^{n-i-1}}{2\, s^n \prod_{i=1}^{n-1} 2i}, \end{aligned} \tag{11}$$

where

$$a_{n,i} = \begin{cases} 0 & (n = i \text{ or } i = -1) \\ 1 & (n = 1 \text{ and } i = 0) \\ a_{n-1,i-1}(n + i - 2) + a_{n-1,i} & (\text{otherwise.}) \end{cases}$$

The resulting value represents the PDF of the summed noise. The expected absolute value of (11) is calculated as

$$E[|x| \mathcal{L}_k(x; s)] = 2 \int_{x=0}^{\infty} x \mathcal{L}_k(x; s) dx = s \prod_{i=1}^{n-1} \frac{2i+1}{2i}. \tag{12}$$

The value of (12) represents the expected magnitude of error compared to the true value. The expected magnitude of error is then adjusted based on the desired value. For example, if the server wants to calculate the final average value, the expected magnitude of error is the value of (12) divided by n. When the target mean absolute error (MAE) of the expected average value is θ, the value of s should be

$$s = n \cdot \frac{\theta \sqrt{\pi} \Gamma(n)}{2\Gamma(1/2 + n)}. \tag{13}$$

The expected squared error is calculated using

$$E[x^2 \mathcal{L}_n(x; s)] = 2\, n s^2. \tag{14}$$

If the server wants to calculate the final average value, the value (14) is divided by n^2. When the target mean squared error (MSE) of the expected average value is θ', the value of s should be

$$s = n \sqrt{\frac{\theta'}{2n}} = \sqrt{\frac{n\theta'}{2}}. \tag{15}$$

Algorithm 1 describes our proposed method.

If the target MSE is desired, Line 3 in Algorithm 1 is replaced by $s \leftarrow \sqrt{n\theta'/2}$.

V. EVALUATION

A. DATASETS

First, we generated synthetic datasets following the normal, uniform, and delta distributions where the values lie in the range [0,100].

Furthermore, four real datasets were evaluated. The first dataset was an email dataset [24]. The first dataset was an

Algorithm 1: Collection and Analysis of LDP Data in a Person-to-Person Interaction.

Input: Δf, target MAE θ or target privacy budget ϵ
Output: Expected average value

/*Process of the data collection server*/
if target MAE is set **then**
  $s \leftarrow \frac{\theta\sqrt{\pi}\Gamma(n)n}{2\Gamma(1/2+n)}$
  $\epsilon' \leftarrow \frac{2\Delta f}{s}$
else if target privacy budget is set **then**
  $\epsilon' \leftarrow \frac{\epsilon}{2}$
end if
Send ϵ' to the users.
/*Process of each user u_i*/
$x'_i \leftarrow x_i + \mathcal{L}(\frac{\Delta f}{\epsilon'})$
Send x'_i to the data collection server.
/*Process of the data collection server*/
$v \leftarrow \frac{1}{n}\sum_{i=1}^{n} x'_i$
Return v.

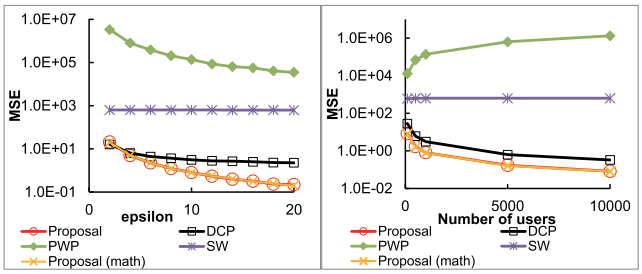
email dataset [27]. We checked the sender, the receiver, and the content of each email in the dataset and identified 19,753 distinct email addresses. Moreover, we counted the number of swear words used by each user. We obtained the list of swear words from https://www.noswearing.com/, which has been used by many studies (e.g., [28], [29], [30]). Information on how many swear words each user emailed was collected under ϵ-LDP.

The second dataset was a who-trusts-whom network dataset [25]. Information on how many users each user trusts was collected under ϵ-LDP. The number of users was 36,692. The minimum and maximum values were 1 and 3,044, respectively.

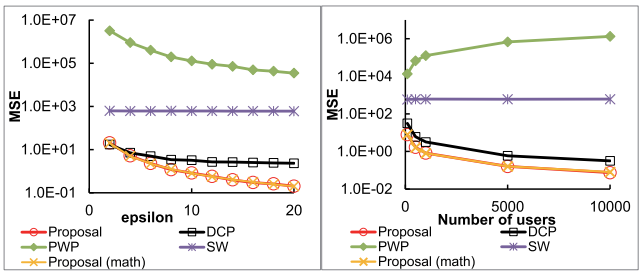
The third dataset included observational contact data from 86 rural Malawian residents [31]. Participants wore pouched sensors on the front of their clothing to detect close proximity. A personal "touch event" between two people was identified when the devices exchanged approximately one radio packet for 20-time intervals. After contact was established, it was considered continuous as long as it kept exchanging no more than one radio packet every second of the subsequent 20-second interval. Each device had an identification number that was used to link contact information established by the individual carrying the device.

The fourth dataset described face-to-face citations from 405 participants at the SFHH conference in Nice, France, which was held on June 4–5, 2009 [32]. Each participant had a device, and at regular intervals, wireless packets were sent with temporary addresses granted to the device. The spatial distance was about 1 m, and the devices were able to detect face-to-face approaches.

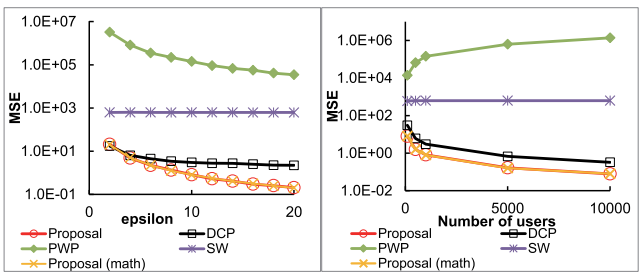
Table 2 summarizes the characteristics of the four datasets.



(a) Normal dist. with varying ϵ (b) Normal dist. with varying n



(c) Uniform dist. with varying ϵ (d) Uniform dist. with varying n



(e) Peak dist. with varying ϵ (f) Peak dist. with varying n

FIGURE 3. Mean squared error (MSE) results of synthetic datasets.

B. EVALUATION RESULTS

We evaluated the effectiveness of our proposed method using synthetic and real datasets. We compared the proposed method with the DCP, PWP, and SW methods proposed in [13], [14] (see Section II). Because these methods do not assume the person-to-person interaction scenario, it is necessary to derive a method for setting the privacy budget value.

For DCP, the maximum ratio of $Pr(\epsilon, x)/Pr(\epsilon, x')$ based on (4) is e^{ϵ} when $x, x' = 1, -1$. In our scenarios, the range of x depending on $x_{i,j}$ is not 2 but $2/(n-1)$. In this case, the maximum ratio is represented by

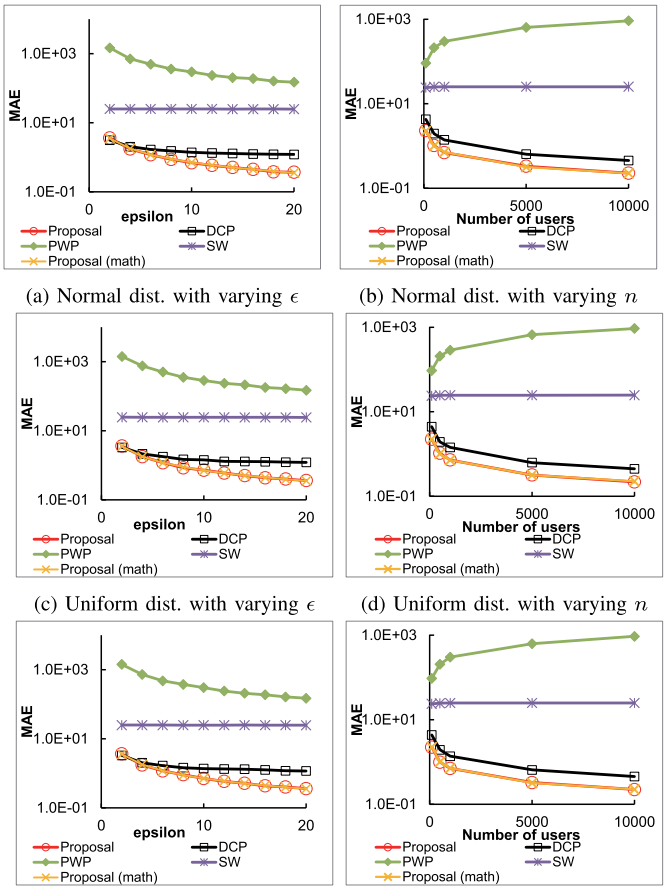
$$\gamma(\epsilon, n) = \frac{Pr(\epsilon, -1 + 2/(n-1))}{Pr(\epsilon, -1)} = \frac{e^{\epsilon} + n - 2}{n - 1}. \quad (16)$$

Therefore, other than u_i, privacy budget $\log \gamma(\epsilon, n)$ is consumed. If the total privacy loss should be ϵ, the privacy budget for x_i should be set to the value obtained using the following equation for ϵ':

$$\epsilon' + (n-1)\log \gamma(\epsilon', n) = \epsilon. \quad (17)$$

TABLE 2. Real Datasets

Database name	Num. of users	Num. of interactions	Min value	Max Value
E-mail dataset [27]	19,753	517,401	0	852
Who-trusts-whom network dataset [25]	75,879	811,480	1	3,044
Village dataset [31]	86	102,293	0	5,398
SFHH dataset [32]	405	70,261	0	2,120

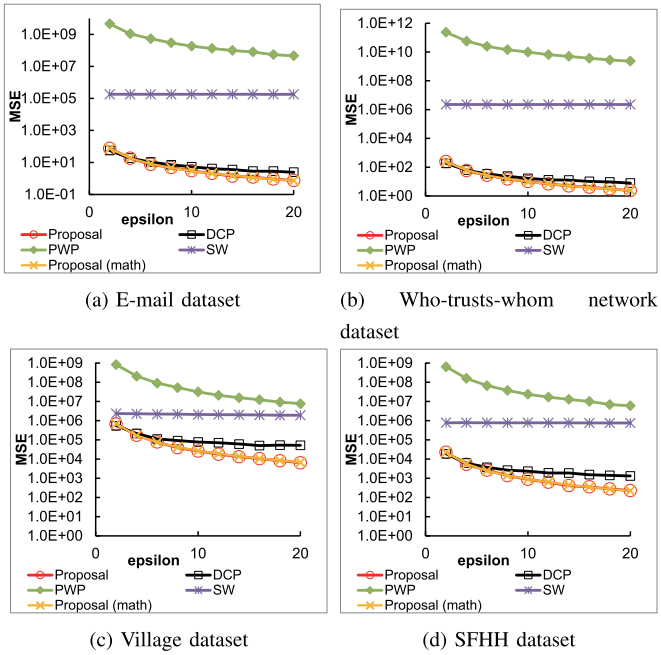


(a) Normal dist. with varying ϵ (b) Normal dist. with varying n

(c) Uniform dist. with varying ϵ (d) Uniform dist. with varying n

(e) Peak dist. with varying ϵ (f) Peak dist. with varying n

FIGURE 4. Mean absolute error (MAE) results of synthetic datasets.



(a) E-mail dataset (b) Who-trusts-whom network dataset

(c) Village dataset (d) SFHH dataset

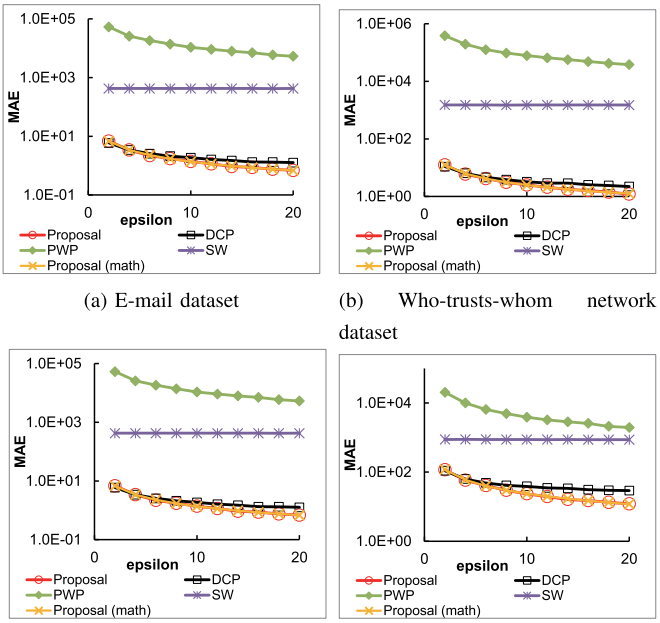
FIGURE 5. MSE results for real datasets. Although the difference between the proposed method and DCP appeared small, the proposed method reduced the MSE by 47%, 40%, 66%, and 62%, respectively, when $\epsilon = 10$.

It is difficult to solve (17) algebraically, but it can be easily solved numerically.

For PWP [13] and SW [14], the consumed privacy budget of u_j is also ϵ when user u_i sends the ϵ-LDP value of x_i to the server. Therefore, when n users send their LDP values to the server, the value of the privacy budget should be ϵ/n to ensure ϵ-LDP.

We experimentally evaluated the MSE and MAE. We repeated each experiment 1,000 times and obtained the average value. The range of ϵ was set to [1, 20] based on [20], [33]. In several existing studies, ϵ was set to smaller values. In a practical case, the range [1, 20] is sufficient for ϵ. In our setting in the synthetic datasets, each true value existed in [0, 100]. When ϵ was 1, the average amount of the Laplace noise per user was 200. The noise was sufficiently large to ensure that the true value was not recognizable at all. When ϵ was 20, the average amount of the Laplace noise per user was 10. Although the privacy protection level was relatively low, this value may be sufficient in some cases. The range of n (number of users) was set to [100, 10000]. The default values of ϵ and n were set to 10 and 1000, respectively.

The MSE results of synthetic datasets are shown in Fig. 3. The results obtained with varying ϵ are shown in Figs. 3(a), (c), and (e) in Fig. 3. The results of *Proposal (math)* represent the mathematical analysis results based on (12) and (14). The results of PWP and SW are worse than those of the other methods. This is because when a user u_i sends x_i value under ϵ'-LDP, this behavior consumes ϵ' of u_i's privacy budget and ϵ' of every u_j's privacy budgets. DCP performed well when ϵ was small. However, for larger ϵ, the proposed method proved to be more effective. Originally, the DCP did not perform well when ϵ was large [20]. Fig. 3(b), (d), and (f) show the results for different numbers of users. As the number of users increases, the amount of noise accumulated increases. However, if the noise added to each value is not too large, they cancel each other out, and the effect of each noise is mitigated. Owing to this tradeoff, the MSE increases or



(a) E-mail dataset (b) Who-trusts-whom network dataset

(c) Village dataset (d) SFHH dataset

FIGURE 6. MAE results for real datasets. Although the difference between the proposed method and DCP appeared small, the proposed method reduced the MAE by 27%, 23%, 41%, and 39%, respectively, when the value of ϵ was 10.

decreases depending on the method. For the proposed method and DCP, the MSE decreased as the number of users increased because each noise was relatively small. In contrast, as PWP had larger noise values, the predicted average MSE increased with the number of users. For all datasets, the results were very similar to each other. As can be seen from (12) and (14), the values of MSE and MAE do not depend on the content of the dataset but the number of users and value of ϵ.

Fig. 4 shows the MAE results. The trend of the results was very similar to that of the MSE results and the MAE values of the proposed method is smaller than those of existing methods.

The experimental results of the MSE are shown in Fig. 5. The performance of DCP and the proposed method were greater than that of other methods for all datasets. It is difficult to read the differences between DCP and the proposed method in Fig. 5, but there are significant differences in the MSE values. When ϵ was 10, the proposed method reduced the MSE by 47%, 40%, 66%, and 62%, respectively, compared with DCP.

Even if the number of noises added to each value is large, the accuracy of the estimation can be increased by collecting a high amount of user data. Therefore, regarding two large datasets (e-mail and who-trusts-whom network datasets), the difference between the proposed method and other methods was relatively small. However, regarding two small datasets (Village and SFHH datasets), it was difficult for all methods to estimate the average value with high accuracy. The proposed method is particularly effective in this difficult task with a small number of users.

The MAE results of real datasets are shown in Fig. 6. The trend of the results is very similar to the MSE results. For all datasets and almost all parameter settings, the proposed method was the most accurate in deriving the average value.

If the server collects data streams from each person, the budget will become small. Therefore, the performances of the proposed method and the DCP are similar in such cases. The performance of the proposed method for small values of ϵ will be improved in the future.

VI. CONCLUSION

In this paper, we defined a novel privacy issue related to person-to-person interactions in LDP. Although this issue has persisted for a long time, it has been overlooked by many organizations. We formalized the problem of LDP in person-to-person interactions and proposed a method that can ensure LDP in such scenarios. Our experimental results showed that the proposed method can reduce the MSE by about 45% compared with the MSE achieved by existing methods. The proposed method is relatively simple; however, the results of this study are only the first step toward solving this important issue. We anticipate that raising this issue will lead to an active discussion.

REFERENCES

[1] S. Ray, T. Palanivel, N. Herman, and Y. Li, "Dynamics in data privacy and sharing economics," *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 114–115, Sep. 2021.

[2] D. Jacobs, T. McDaniel, A. Varsani, R. U. Halden, S. Forrest, and H. Lee, "Wastewater monitoring raises privacy and ethical considerations," *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 116–121, Sep. 2021.

[3] Differential-Privacy-Team, "Learning with privacy at scale," 2016. [Online]. Available: https://machinelearning.apple.com/research/learning-with-privacy-at-scale

[4] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.

[5] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 3574–3583.

[6] F. Aldà and H. U. Simon, "A lower bound on the release of differentially private integer partitions," *Inf. Process. Lett.*, vol. 129, pp. 1–4, Jan. 2018.

[7] 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, 2013.

[8] Y. Sei, J. A. Onesimu, H. Okumura, and A. Ohsuga, "Privacy-preserving collaborative data collection and analysis with many missing values," *IEEE Trans. Dependable Secure Comput.*, early access, May 13, 2022, doi: 10.1109/TDSC.2022.3174887.

[9] G. Cormode, S. Maddock, and C. Maple, "Frequency estimation under local differential privacy," in *Proc. VLDB*, 2021, pp. 2046–2058.

[10] I. Mironov, "Rényi differential privacy," in *Proc. IEEE Comput. Secur. Foundations Symp.*, 2017, pp. 263–275.

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

[12] G. Varma, R. Chauhan, and D. Singh, "Sarve: Synthetic data and local differential privacy for private frequency estimation," *Cybersecurity*, vol. 5, no. 1, pp. 1–20, Dec. 2022. [Online]. Available: https://link.springer.com/articles/10.1186/s42400-022-00129-6

[13] Q. Xue, Y. Zhu, and J. Wang, "Mean estimation over numeric data with personalized local differential privacy," *Front. Comput. Sci.*, vol. 16, no. 3, pp. 1–10, 2022.

[14] 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*, 2020, pp. 621–635.

[15] H. Navidan, V. Moghtadaiee, N. Nazaran, and M. Alishahi, "Hide me behind the noise: Local differential privacy for indoor location privacy," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops*, 2022, pp. 514–523.
[16] J. W. Kim and B. Jang, "Workload-aware indoor positioning data collection via local differential privacy," *IEEE Commun. Lett.*, vol. 23, no. 8, pp. 1352–1356, Aug. 2019.
[17] X. Ren, L. Shi, W. Yu, S. Yang, C. Zhao, and Z. Xu, "LDP-IDS: Local differential privacy for infinite data streams," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2022, vol. 14, pp. 1064–1077.
[18] X. Ren et al., "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.
[19] Y. Zhao et al., "Local differential privacy-based federated learning for Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8836–8853, Jun. 2021.
[20] L. Cui, J. Ma, Y. Zhou, and S. Yu, "Boosting accuracy of differentially private federated learning in industrial IoT with sparse responses," *IEEE Trans. Ind. Inform.*, vol. 19, no. 1, pp. 910–920, Jan. 2023.
[21] R. Hu, Y. Guo, and Y. Gong, "Concentrated differentially private federated learning with performance analysis," *IEEE Open J. Comput. Soc.*, vol. 2, pp. 276–289, Jul. 2021.
[22] S. Wang and J. M. Chang, "Privacy-preserving boosting in the local setting," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4451–4465, 2021.
[23] Y. Sei, H. Okumura, and A. Ohsuga, "Re-identification in differentially private incomplete datasets," *IEEE Open J. Comput. Soc.*, vol. 3, pp. 62–72, 2022.
[24] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, 2009.
[25] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," in *Proc. Int. Semantic Web Conf.*, 2003, pp. 351–368.
[26] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations Trends Theor. Comput. Sci.*, vol. 9, no. 3/4, pp. 211–407, 2014.
[27] W. W. Cohen, "Enron email dataset," 2015. [Online]. Available: https://www.cs.cmu.edu/∼./enron/
[28] E. W. Pamungkas, V. Basile, and V. Patti, "Do you really want to hurt me? predicting abusive swearing in social media," in *Proc.12th Lang. Resour. Eval. Conf.*, 2020, pp. 6237–6246.
[29] L. P. V. Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in *Human-Inspired Computing and its Applications*, vol. 8856. Berlin, Germany: Springer, 2014, pp. 221–232. [Online]. Available: https://link.springer.com/chapter10.1007/978-3-319-13647-9_21
[30] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and techniques," in *Proc. Annu. ACM Web Sci. Conf.*, 2013, pp. 195–204. [Online]. Available: www.noswearing.com
[31] L. Ozella et al., "Using wearable proximity sensors to characterize social contact patterns in a village of rural Malawi," *EPJ Data Sci.*, vol. 10, no. 1, pp. 1–17, Sep. 2021. [Online]. Available: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00302-w
[32] M. Génois and A. Barrat, "Can co-location be used as a proxy for face-to-face contacts?," *EPJ Data Sci.*, vol. 7, no. 1, pp. 1–18, May 2018. [Online]. Available: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0140-1
[33] Y. Wang, Y. Tong, and D. Shi, "Federated latent dirichlet allocation: A local differential privacy based framework," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6283–6290.

YUICHI SEI (Member, IEEE) received the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2009. From 2009 to 2012, he was with the Mitsubishi Research Institute, Tokyo. He joined The University of Electro-Communications, Chofu, Japan, in 2013. He is currently an Associate Professor with the Graduate School of Informatics and Engineering, Chofu. He is also a Fellow Researcher with Mitsubishi Research Institute and an Adjunct Researcher with Waseda University, Tokyo. His research interests include agent computing, privacy-preserving data mining, and software engineering. He was the recipient of IPSJ/IEEE Computer Society Young Computer Researcher Award in 2021.

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