TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 16/18 pp412–421
DOI: 10.26599/TST.2021.9010027
Volume 27, Number 2, April 2022

PrivBV: Distance-Aware Encoding for Distributed Data with Local Differential Privacy

Lin Sun, Guolou Ping, and Xiaojun Ye*

Abstract: Recently, local differential privacy (LDP) has been used as the de facto standard for data sharing and analyzing with high-level privacy guarantees. Existing LDP-based mechanisms mainly focus on learning statistical information about the entire population from sensitive data. For the first time in the literature, we use LDP for distance estimation between distributed data to support more complicated data analysis. Specifically, we propose PrivBV—a locally differentially private bit vector mechanism with a distance-aware property in the anonymized space. We also present an optimization strategy for reducing privacy leakage in the high-dimensional space. The distance-aware property of PrivBV brings new insights into complicated data analysis in distributed environments. As study cases, we show the feasibility of applying PrivBV to privacy-preserving record linkage and non-interactive clustering. Theoretical analysis and experimental results demonstrate the effectiveness of the proposed scheme.

Key words: local differential privacy; privacy-preserving data publishing; non-interactive clustering

1 Introduction

In recent years, a lot of data have been crowdsourced and comprehended continuously for decision-making in many data-driven applications[1]. However, in many applications, personal data are aggregated to improve user experiences and the quality of service, which leads to potential privacy leakage. For example, the browsing records used in a recommender system could reveal personal shopping habits. Thus, the need to protect individual privacy has become a major public concern in privacy-preserving data publishing[2].

Recently, a variety of encoding mechanisms that claim to be privacy-preserving have been proposed in specific application fields. For example, in healthcare record linkage, anonymization methods that support similarity comparisons, such as the Bloom Filter-based mechanism (NBF[3]), Low-Cost Bloom Filters (LCBF[4]), and Bit Vector (BV[5]), are used for numerical data encoding. In the preservation of location information[6], several anonymity-based mechanisms and transformation-based approaches have been proposed to protect users' exact locations while providing desired location-based services. However, these proposed methodologies lack strict privacy-guarantee models and might be vulnerable to background knowledge attacks.

The notion of local differential privacy (LDP[7]) has been deemed a de facto standard for privacy-preserving data analysis since its proposal. Without loss of generality, existing LDP protocols can be specified into three steps. First, each distributed user encodes his data into a specific data type. Then, the encoded data is perturbed to generate noisy data. Finally, perturbed data from distributed users are aggregated by the data collector to deploy data analysis tasks. The ϵ-LDP provides high privacy-preserving levels by guaranteeing that the probability of any two different inputs being projected into the same output is bounded by e^{ϵ}. LDP has shown great significance in learning statistical information, especially for the mean and frequency estimations (heavy hitter estimation)[8–11].

However, two major challenges remain in applying LDP to real-life applications. First, existing mechanisms

• Lin Sun, Guolou Ping, and Xiaojun Ye are with School of Software, Tsinghua University, Beijing 100084, China. E-mail: sunl16@mails.tsinghua.edu.cn; pgl19@mails.tsinghua.edu.cn; yexj@tsinghua.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2020-10-26; accepted: 2021-03-24

with LDP guarantees can only estimate statistical information from large volumes of data. Data utilities are of limited scope in the perturbed space because the randomization introduces too much noise in the perturbation process. To preserve privacy, data from the user are randomized to discrete space, which distorts the data estimation. For example, the 1BitMean mechanism[12] randomizes a numerical value into one bit for mean estimation. For the aggregator, estimating population means causes large errors under a limited amount of data. Second, commonly used perturbing mechanisms randomize the users' value in the original space, which causes a sharp privacy leakage in private data sharing when the privacy budget ϵ is set at a high level. For example, the Laplace mechanism[13] and the Piecewise mechanism[14] add a random noise located near 0 with a high probability when the privacy budget is high. The range of original data can be inferred from the perturbed data with high confidence.

To improve data utilities in the perturbed space (even when the privacy budget is not low), we consider an anonymization-perturbation solution to achieve (ϵ,δ)-LDP in this paper. More specifically, we take the Bit Vector[5] mechanism as the basic anonymization and expand it to be locally differentially private (PrivBV) by applying the randomized response. We validate that the similarity between two records (by Euclidean distance) can be preserved in the perturbation procedure. With this distance-aware property in the anonymized space, this mechanism can be further used for privacy-preserving data publishing and many complicated analyzing tasks. As far as we know, no solution has been proposed for non-interactive clustering under LDP. As an improvement, we also propose a clustering algorithm that only relies on the distance information between different records.

Generally, the contributions of this paper can be summarized as follows. (1) We improve the capabilities of the Bit Vector mechanism by discovering the distance-continuation property in the anonymized space to make it capable of the entire range distance estimation. (1) We propose PrivBV: an (ϵ,δ)-LDP mechanism with a distance-aware property. Then, we discover an optimization method to achieve a lower estimation error for high-dimensional data. (3) We explore the application area of the proposed PrivBV mechanism. For the first time, we present one possible solution to non-interactive clustering and show the feasibility of applying PrivBV in privacy-preserving record linkage.

The overall structure of this paper can be divided into several parts. In Section 2, we present related work and some preliminaries. In Section 3, we first propose a distance-continuation algorithm for the entire range distance estimation of BV. Then, we propose PrivBV and an optimization strategy for applying PrivBV on high-dimensional data. As a study case, we delineate the clustering algorithm in Section 4. The experimental results are analyzed in Section 5. Finally, we conclude this paper and discuss future work in Section 6.

2 Preliminaries

Before introducing several fundamental mechanisms that are included in our proposed solution, definitions of the notations are listed in Table 1. For convenience, each dimension is assumed to be in the Euclidean space.

2.1 Distance-aware encoding mechanism

The distance-aware encoding schemes try to encode data into an anonymized space and preserve the Euclidean distance. In the application of privacy-preserving record linkage, researchers combine Bloom Filters and N-gram to detect string similarities with a privacy guarantee[3]. As an improvement of the Bloom Filter-based solutions, the Bit Vector (BV[4, 5]) mechanism is proposed to retrieve the Euclidean distance in the anonymized space. Originally, BV is used for numerical data linking in the three-party model[15].

The BV mechanism requires a set of random values $\mathcal{R} = \{r_1, r_2, ..., r_s\}$ and interval parameter t. Each random value follows a uniform distribution in the input domain $\mathcal{X} = [L, U]$ with $\mu = U - L$. These parameters can be transferred through secure channels or using encryption when deployed in real-world applications. On the basis of $\mathcal{R}$ and t, a hash family is defined by

$$H = \{h_i\}, h_i(x) = \begin{cases} 1, & \text{if } x \in [r_i - t, r_i + t]; \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Equation (1) encodes a numerical value into a bit vector with s components $B_{i:i\in[s]}$, which is composed of several set components (with $B_i = 1$) and unset

Table 1 Notations.

Notation	Explanation
$\mathcal{R}, t, s$	Parameters of BV mechanism
$[L, U]$	Domain of source data, and $\mu = U - L$
$\mathcal{B}^x$	Bit vector of data x
$d_E(\cdot), d_H(\cdot)$	Euclidean (Hamming) distance
$\hat{d}_E(\cdot), \hat{d}_H(\cdot)$	Estimated Euclidean (Hamming) distance
ϵ, δ	Privacy budget defined by LDP

components (with $B_i = 0$). BV has two important features. First, the number of set components is identical for different scalar values in expectation. Thus, BV provides indistinguishabilities. Moreover, the Euclidean distance can be estimated in the anonymized Hamming space. For values x_1, x_2 with $d_E = |x_1 - x_2|$, the distance can be estimated as $\hat{d_E} = d_H \cdot \mu/(2s)$. With these two properties, BV can be used for privacy-preserving distance estimation applications.

Despite its efficacy, BV still suffers from two major disadvantages, i.e., data utility insufficiency and possible privacy leakage. On the one hand, the BV mechanism can only estimate the distance with $d_E \leqslant 2t$ (detailed in Section 3.1). This property leads to a high estimation error when the interval parameter t is incorrectly set. On the other hand, knowing the set of random values $\mathcal{R}$, the range of original data can typically be narrowed, which violates the privacy-preservation demand in the data sharing process. For example, knowing that $r_1 = 6.3, r_2 = 5.7$ when $t = 1$, the encoded result BV $(x) = (1, 0)$ indicates that $x \in [5.3, 7, 3]$ and $x \notin [4.7, 6.7]$. Thus, $x \in [6.7, 7.3]$ by inference. With more background knowledge of random values and t, the range of x can be further determined.

2.2 Local differential privacy

The LDP[7, 9, 14, 16] is proposed as an extension of differential privacy (DP[17]) to provide strong privacy guarantees in the local context. In the local setting, an untrusted aggregator wants to learn statistical information from users. For privacy concerns, every user encodes and perturbs his data with an LDP mechanism $\mathcal{M}$ and shares the randomized data with the aggregator. Mechanisms with LDP guarantees are defined as follows:

Definition 1 ((ϵ, δ)-LDP) A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-LDP iff for all $\mathcal{S} \subset \text{Range}(\mathcal{M})$ and all x_1 and x_2 in the input domain:

$$\Pr[\mathcal{M}(x_1) \in \mathcal{S}] \leqslant \mathrm{e}^{\epsilon} \cdot \Pr[\mathcal{M}(x_2) \in \mathcal{S}] + \delta \tag{2}$$

Intuitively, (ϵ, δ)-LDP means that with a probability of at least $1 - \delta$, the input tuple cannot be distinguished when the aggregator receives the output S. The LDP has been broadly used in statistical aggregation, such as mean[18] and frequency[19] estimation. In addition, it has been deployed in many well-known systems. Google proposed RAPPOR[20] to study clients behavior, such as surfing habits in Chrome. In Window Insiders, the 1BitMean mechanism[12] is used for mean estimation.

One inherent problem of mechanisms with LDP is that the amount of data plays an essential role in balancing data utilities and privacy guarantees. The ϵ-LDP implies that the data utility in a single record is negligible. This implication raises one research interest: can the perturbed data be compared with each other in the local mode? In this paper, we provide new insights on improving data utility in (ϵ, δ)-LDP by expanding the perturbed probability space.

2.3 Problem definition and evaluation matrix

This paper studies the problem of distributed data aggregation and analysis in the local setting of differential privacy. To preserve a privacy guarantee, each scalar value v_i from distributed users is perturbed by a privacy-preserving encoding mechanism $\mathcal{M}(v_i)$. Then, an untrusted aggregator collects the encoded data and wants to analyze data from the user's side, including the distance estimation and clustering.

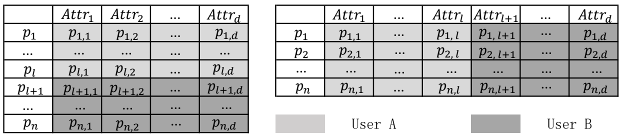
Privacy-preserving distance estimation. For one-dimensional data $x, y \in \mathbb{R}$ and the perturbed data by a privacy-preserving mechanism $\mathcal{M}$, the Euclidean distance $d_E = |x - y|$ is estimated in the perturbed space as $\hat{d_E} = \mathcal{D}(\mathcal{M}(x), \mathcal{M}(y))$. This paper also considers distance estimation over d-dimensional data. The Euclidean distance over $x = [x_1, x_2, ..., x_d]$ and $y = [y_1, y_2, ..., y_d]$ is $d_E = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$. As shown in Fig. 1, data from different users share the same attribute in the horizontal settings, while different users have different attributes in the vertical setting. The error of distance estimation is enhanced by the average absolute error $E[|\hat{d_E} - d_E|]$.

Privacy-preserving record linkage. PPRL aims to find matched records across data sources D_A and D_B. Given threshold T, the record pair (x, y) from different data sources is classified as matches iff $\mathcal{D}(x, y) \leqslant T$. Otherwise, (x, y) is non-matches. The performance of PPRL is evaluated by the precision (P), recall (R), and F1-score, defined as

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{F1} = \frac{2PR}{P + R},$$

where TP, FP, and FN are the number of true matches, false matches, and false non-matches, respectively.

Non-interactive clustering. Given n distributed



	$Attr_1$	$Attr_2$	...	$Attr_d$
p_1	$p_{1,1}$	$p_{1,2}$	...	$p_{1,d}$
...	...	...	...	...
p_l	$p_{l,1}$	$p_{l,2}$	...	$p_{l,d}$
p_{l+1}	$p_{l+1,1}$	$p_{l+1,2}$	...	$p_{l+1,d}$
...	...	...	...	...
p_n	$p_{n,1}$	$p_{n,2}$	...	$p_{n,d}$

	$Attr_1$	...	$Attr_l$	$Attr_{l+1}$	...	$Attr_d$
p_1	$p_{1,1}$	...	$p_{1,l}$	$p_{1,l+1}$	...	$p_{1,d}$
p_2	$p_{2,1}$	...	$p_{2,l}$	$p_{2,l+1}$	...	$p_{2,d}$
...	...	...	...	...	...	...
p_n	$p_{n,1}$	...	$p_{n,l}$	$p_{n,l+1}$	...	$p_{n,d}$

User A
User B

Fig. 1 Horizontally and vertically partitioned data.

users in which each user owns one record, the non-interactive clustering task aims at grouping the user records into k groups without interactively communicating with users. One essential challenge in the non-interactive clustering with privacy guarantees is how to calculate similarities between different records in the anonymized space. The clustering performance is measured by the normalized mutual Information[21].

3 Improving Data Utilities and Privacy-Preserving Levels of BV

As described in Section 2, BV suffers from data utility insufficiency and possible privacy leakage. To improve the performance of BV, we define the property of distance-continuation and show that greater distances can be adjusted to reduce the estimation error. Then, we propose the PrivBV—a distance-aware perturbation mechanism under (ϵ,δ)-LDP based on the BV mechanism. Furthermore, we present an optimization strategy for applying PrivBV in high-dimensional data.

3.1 Distance-continuation

Figure 2 shows the distance estimation result by the BV mechanism. We can learn from the example of (y_2, x_6) that when $d_E(\mathcal{B}^x, \mathcal{B}^y) > 2t$, the distance estimation utilities lose. This is because the information of value x is preserved by the range $[x-t, x+t]$ when encoding. We will show that we can still recover the distance information in the perturbed space.

As we noted, the data are continuous in the domain of $\mathbb{R}$, which is for $x < y < z$, and we have $d_E(x,z) = d_E(x,y) + d_E(y,z)$. In the anonymized space of BV, the continuation property is still preserved.

Theorem 1 (Distance-continuation of the BV) For numerical values $x, y, z \in \mathbb{R}$ with $x < y < z$ and $d_E(x,y), d_E(x,z), d_E(y,z) \leqslant 2t$, we have

$$\hat{d}_E(x,z) = \hat{d}_E(x,y) + \hat{d}_E(y,z).$$

Proof Let $\tau = \hat{d}_E(x,y) + \hat{d}_E(y,z) - \hat{d}_E(x,z)$. It then holds that:

$$\tau = \frac{u}{2s} \cdot [d_H(\mathcal{B}^x, \mathcal{B}^y) + d_H(\mathcal{B}^y, \mathcal{B}^z) - d_H(\mathcal{B}^x, \mathcal{B}^z)] = \frac{u}{2s} \cdot \sum_{i=1}^{s} \left[d_H(\mathcal{B}_i^x, \mathcal{B}_i^y) + d_H(\mathcal{B}_i^y, \mathcal{B}_i^z) - d_H(\mathcal{B}_i^x, \mathcal{B}_i^z)\right],$$

where $\mathcal{B}^x, \mathcal{B}^y$, and $\mathcal{B}^z$ are the bit vector of x, y, and z by the PrivBV mechanism. For simplicity, let $\iota \in \{0,1\}$ represent the value of the i-th bit. As an example, when $\iota = 0$, the triple $[\iota, \iota, \bar{\iota}]$ is $\left[\mathcal{B}_i^x, \mathcal{B}_i^y, \mathcal{B}_i^z\right] = [0,0,1]$ and the Hamming distances of the i-th bit are $d_H(\mathcal{B}_i^x, \mathcal{B}_i^y) = 0, d_H(\mathcal{B}_i^y, \mathcal{B}_i^z) = 1$, and $d_H(\mathcal{B}_i^x, \mathcal{B}_i^z) = 1$.

In the anonymized space, for the i-th bit, all possible situations of $\mathcal{B}_i^x, \mathcal{B}_i^y$, and $\mathcal{B}_i^z$ are listed in the following table.

As shown in Table 2, situation ③ violates the continuation property. Without loss of generality, we assume that $\tau = 0$, corresponding to $[\mathcal{B}_i^x, \mathcal{B}_i^y, \mathcal{B}_i^z] = [0,1,0]$. According to the hash function by Eq. (1), there exists a random value r_i, such that:

$$\begin{cases} x \in [L, r_i - t]; \\ y \in [r_i - t, r_i + t]; \\ z \in [r_i + t, U] \end{cases} \tag{3}$$

Equation (3) means that $(z-t)-(x+t) = (z-x)-2t > 0$, which does not hold because $z - x \leqslant 2t$. Thus the case of $\tau = 0$ is invalid. Analogously, when $\tau = 1$, the situation of $[\mathcal{B}_i^x, \mathcal{B}_i^y, \mathcal{B}_i^z] = [1,0,1]$ is also unsatisfied. The distance-continuation property in the anonymized space is guaranteed. ■

	$x_1=4$	$x_2=5$	$x_3=6$	$x_4=7$	$x_5=8$	$x_6=9$
$y_1=4$	0	1.015	1.975	2.51	2.585	2.585
$y_2=5$	1.015	0	0.96	2.025	2.64	2.64
$y_3=6$	1.975	0.96	0	1.065	2.07	2.62
$y_4=7$	2.51	2.025	1.065	0	1.005	2.035
$y_5=8$	2.585	2.64	2.07	1.005	0	1.03
$y_6=9$	2.585	2.64	2.62	2.035	1.03	0

Fig. 2 Example of distance estimation by BV. The estimated distance with a gray background is inaccurate. We set $t=1.2$ and $s=1000$. The range of original data is $[0,20]$.

With the distance-continuation property, we can use the truly estimated distances to optimize the wrong estimations. Following Fig. 2, we can adjust $\hat{d}_E(5,8) = \hat{d}_E(5,7) + \hat{d}_E(7,8) = 3.03$. The distance-continuation also guarantees the uniqueness of the estimated distance. For example, the result of $\hat{d}_E(5,7) + \hat{d}_E(7,8)$ is identical to that of $\hat{d}_E(5,6) + \hat{d}_E(6,8)$.

Currently, distances over $2t$ cannot be adjusted without those records in $2t$ as springboards. To solve this problem, we recommend adding intermediate values when encoding. For example, when $2t < 3$, the distance between 3.4 and 7.9 cannot be estimated. The data owner can then generate a noisy value $v = 5.5$. In this way, the Euclidean distance can be estimated as

Table 2 Possible values of bit vectors.

Situation	$\mathcal{B}_i^x$	$\mathcal{B}_i^y$	$\mathcal{B}_i^z$	$\tau = 0$
①	ι	ι	ι	✓
②	ι	ι	$\bar{\iota}$	✓
③	ι	$\bar{\iota}$	ι	×
④	ι	$\bar{\iota}$	$\bar{\iota}$	✓

$\hat{d}_E(3.4, 5.5) + \hat{d}_E(5.5, 7.9)$. Notably, adding external values increases computing complexity.

The pseudocode of the distance-adjusting algorithm is described in Algorithm 1. Given n records in the anonymized space encoded by the BV mechanism, we first estimate the Euclidean distance $D_{i,j}$ using the BV mechanism (Lines 1 and 2). As the distance-continuation only holds within the range of $2t$, we then need to estimate $2t$ by finding the maximum domain that has the distance-continuation property (Line 3). While adjusting the distance matrix D, we first keep distances within T unchanged (Line 5). Then, distances over T are updated using Theorem 1 (Lines 6–8). Finally, the remaining distances are kept unchanged (Line 9). In our implementation, a flag matrix recording in which iterations $D_{i,j}$ are revised is included, and the distance can only be updated with a modified distance before the current iterations.

3.2 PrivBV

Given a random value r_i, the probability function of BV can be written as $\Pr[\mathcal{B}_i = 1 | r_i, t] = \Pr[x \in [r_i - t, r_i + t]]$. As we showed in the description of the BV mechanism, once the random values are known, the range of original data can be guessed with high confidence. This circumstance indicates that BV mechanism is vulnerable under the observation of r_i. In this section, we will extend the BV mechanism to be (ϵ, δ) locally differentially private. Although the random values $\mathcal{R}$ are known by the aggregator by chance, privacy can still be preserved under LDP.

For data utility purposes, we want the extended mechanism to be distance-aware. BV can preserve distance information in the Hamming space, because for numerical data x_1, x_2, the probability of whether x_1 and x_2 remain in the same interval is proportional to $d_E(x_1, x_2)$. Inspired by the randomized response, we proposed PrivBV by designing a 1Bit-like mechanism in each set bit. The probability function of PrivBV is (due to the space consideration, we use $C_\epsilon = \dfrac{e^\epsilon + 1}{e^\epsilon - 1}$ throughout this paper):

$$\Pr[\mathcal{B}_i = 1 | r_i, t] = \frac{\Pr[x \in [r_i - t, r_i + t]]}{C_\epsilon} + \frac{1}{e^\epsilon + 1}.$$

Algorithm 1 Distance-adjusting by Distance-Continuation

```
Input: $n$ distributed users and each user owns vector $v_i$, where
    $v_i$ is the encoded result by bit vector with $\mu$ and $s$.
Output: the adjusted distance matrix $\hat{D}$
1: Generate a distance matrix $D \in \mathbb{R}^{n \times n}$.
2: Estimate the distance $D_{i,j}$ by BV mechanism
       $\forall i, j \in [n] : D_{i,j} = d_H(v_i, v_j) \cdot \mu/(2s)$.
3: Estimating the range of $2t$ by
       $T = \max\{D_{i,k} | \exists i, j, k : D_{i,j} + D_{j,k} = D_{i,k}\}$.
4: initialize a new distance matrix $\hat{D}$, $\forall i, j, \hat{D}_{i,j} = \infty$.
5: $\forall D_{i,j} \leqslant T$: $\hat{D}_{i,j} = D_{i,j}$.
6: for $\hat{D}_{i,j} \leqslant T$ and $\hat{D}_{j,k} \leqslant T$ and $\hat{D}_{i,k} = \infty$ do
7:     update $\hat{D}_{i,k} = \hat{D}_{i,j} + \hat{D}_{j,k}$.
8: end for
9: $\forall \hat{D}_{i,j} = \infty$, $\hat{D}_{i,j} = D_{i,j}$.
10: return $\hat{D}$.
```

Theorem 2 (Expected number of set components) In PrivBV, the expected number of components that are set in each vector is

$$E[w] = s \cdot \left(\frac{2t}{\mu \cdot C_\epsilon} + \frac{1}{e^\epsilon + 1}\right).$$

Theorem 2 indicates that in the anonymized space of PrivBV, the expected number of components for different source values remains the same, which provides indistinguishability for the original values. Furthermore, with the use of the randomized response, PrivBV achieves (ϵ, δ)-LDP.

Theorem 3 Given s random values $r_1, r_2, ..., r_s$, PrivBV satisfies (ϵ, δ)-LDP, where

$$\delta = \left(\frac{e^\epsilon}{e^\epsilon + 1}\right)^s - e^\epsilon \cdot \left(\frac{1}{e^\epsilon + 1}\right)^s.$$

Proof Let x be numerical data in the input domain, B be the bit vector encoded by BV, and $\mathcal{B}$ be the randomized vector by PrivBV. We then have:

$$\Pr[\mathcal{B}|x] = \Pr[\mathcal{B}|B, x] \cdot \Pr[B|x] = \Pr[\mathcal{B}|B].$$

The above equation turns $\Pr[\mathcal{B}|x]$ into $\Pr[\mathcal{B}|B]$. We consider the situation for encoding with one bit ($s = 1$). Let b_i be the i-th bit in b. Without loss of generality, for the i-th bit in $\mathcal{B}_i$ and B_i, PrivBV indicates that

$$\Pr[\mathcal{B}_i | B_i] = \left(\frac{e^\epsilon}{e^\epsilon + 1}\right)^{\mathcal{B}_i \odot B_i} \cdot \left(\frac{1}{e^\epsilon + 1}\right)^{1 - \mathcal{B}_i \odot B_i}.$$

where the $\mathcal{B}_i \odot B_i$ operation returns 1 if $\mathcal{B}_i = B_i$, and 0 otherwise. Considering all random values $r_1, r_2, ..., r_s$, for any x in the input domain and $\mathcal{B}$ in the output domain, we have:

$$\Pr[\mathcal{M}(x) = \mathcal{B}] = \left(\frac{e^\epsilon}{e^\epsilon + 1}\right)^{\mathcal{B}_1 \odot b_1} \cdot \left(\frac{1}{e^\epsilon + 1}\right)^{1 - \mathcal{B}_1 \odot b_1} \times \cdots \times \left(\frac{e^\epsilon}{e^\epsilon + 1}\right)^{\mathcal{B}_s \odot b_s} \cdot \left(\frac{1}{e^\epsilon + 1}\right)^{1 - \mathcal{B}_s \odot b_s} = \left(\frac{e^\epsilon}{e^\epsilon + 1}\right)^{\sum_{i=1}^{s} (\mathcal{B}_i \odot b_i)} \cdot \left(\frac{1}{e^\epsilon + 1}\right)^{s - \sum_{i=1}^{s} (\mathcal{B}_i \odot b_i)},$$

where $\mathcal{M}(\cdot)$ denotes the PrivBV mechanism. Thus, it holds that $\Pr[\mathcal{B}|B] \leqslant \left(\frac{e^{\epsilon}}{e^{\epsilon}+1}\right)^s$, which means for any input pair (x_1, x_2) and output $\mathcal{B}$, the PrivBV mechanism guarantees:

$$\Pr[\mathcal{M}(x_1)=\mathcal{B}] \leqslant e^{\epsilon}\cdot\Pr[\mathcal{M}(x_2)=\mathcal{B}]+\delta.$$

According to the definition of LDP, PrivBV satisfies (ϵ,δ)-LDP, where $\delta = \left(\frac{e^{\epsilon}}{e^{\epsilon}+1}\right)^s - e^{\epsilon}\cdot\left(\frac{1}{e^{\epsilon}+1}\right)^s$. ■

Theorem 3 gives the bound of the privacy-preserving level. When used in applications, s is usually very large. Taking $s=1000$ as an example, the PrivBV mechanism satisfies $(2, 7.5\times10^{-56})$-LDP. The PrivBV encoding mechanism guarantees privacy for original data. Now, we focus on retrieving the Euclidean distance from the Hamming space. We first show that distances in Hamming space are correlated to Euclidean distances.

Theorem 4 For values x_1, x_2 in the input domain and the Hamming distance d_H between encoded vectors, e can estimate d_E as

$$\hat{d}_E = \frac{\mu\cdot C_{\epsilon}^2}{2s}\cdot d_H - \frac{\mu\cdot e^{\epsilon}}{(e^{\epsilon}-1)^2}.$$

Proof As stated in the PrivBV mechanism, the process of applying randomization to the BV mechanism can be considered flipping the bits in BV with probability $\frac{1}{e^{\epsilon}+1}$. Thus, the expected Hamming distance can be estimated as

$$E[d_H] = 2s\frac{d_E}{\mu}\frac{e^{2\epsilon}+1}{(e^{\epsilon}+1)^2} + \left(s-2s\frac{d_E}{\mu}\right)\frac{2e^{\epsilon}}{(e^{\epsilon}+1)^2}.$$

Thus, d_H in the perturbed space can be used to estimate the Euclidean distance d_E. ■

Theorem 4 shows why the proposed PrivBV can estimate the Euclidean distance. We now give the upper bound of the estimation error in the following theorem.

Theorem 5 The estimation error of PrivBV is

$$|\hat{d}_E - d_E| = \mathcal{O}\left(\frac{\mu\cdot C_{\epsilon}^2}{2}\cdot\sqrt{\frac{\ln\frac{2}{\beta}}{2s}}\right).$$

Proof According to the Chernoff-Hoeffding bound[12], we have:

$$\Pr\left[|d_H - E[d_H]| \geqslant t\right] \leqslant 2\cdot e^{-\frac{2t^2}{s}}.$$

Considering the relationship between d_H and d_E, it then holds that

$$\Pr\left[\left|\frac{2s}{C_{\epsilon}^2}\cdot\frac{\hat{d}_E}{\mu} - \frac{2s}{C_{\epsilon}^2}\cdot\frac{d_E}{\mu}\right| \geqslant t\right] \leqslant 2\cdot e^{-\frac{2t^2}{s}},$$

which equals

$$\Pr\left[\left|\hat{d}_E - d_E\right| \geqslant \frac{\mu t\cdot C_{\epsilon}^2}{2s}\right] \leqslant 2\cdot e^{-\frac{2t^2}{s}} \qquad (4)$$

Thus, by setting $t=\theta\cdot 2s/C_{\epsilon}^2$, we obtain:

$$\Pr[|\hat{d}_E - d_E| \geqslant \theta\mu] \leqslant 2\cdot e^{-8s\theta^2/C_{\epsilon}^4}.$$

Finally, letting $\beta = 2\cdot e^{-8s\theta^2/C_{\epsilon}^4}$, the estimation error is

$$\theta\mu \leqslant \frac{\mu}{2}\cdot C_{\epsilon}^2\cdot\sqrt{\frac{\ln\frac{2}{\beta}}{2s}}.$$

Thus, the error of PrivBV is bounded by a probability of at least $1-\beta$. ■

3.3 Optimization of high-dimensional data

The previous section focused on applying PrivBV for one-dimensional data analysis. In real-world applications, data collected from different sources are usually multi-dimensional. As shown in Fig. 1, data from the user's side can be horizontally and vertically partitioned. In horizontal and vertical settings, given record pair $p_A, p_B \in \mathbb{R}^d$, the goal of distance estimation is

$$d_E(p_A, p_B) = \sqrt{\sum_{i=1}^{d}(p_{A,i}-p_{B,i})^2} \qquad (5)$$

where $p_{A,i}$ and $p_{B,i}$ are the i-th attribute of records p_A and p_B, respectively.

The substantive solution to Eq. (5) is obtained by encoding each dimension independently and estimating the distance between records as

$$\hat{d}_E = \frac{\mu C_{\epsilon}^2}{2s}\cdot\sqrt{\sum_{i=1}^{d}\left(d_H(p_{A,i}, p_{B,i}) - \frac{2s\cdot e^{\epsilon}}{(e^{\epsilon}+1)^2}\right)^2}.$$

For distributed data custodians with vertically partitioned data, distance estimation can be improved. In Fig. 1, we define $L_{A,B}^2 = \sum_{i=1}^{l}(p_{A,i}-p_{B,i})^2$, $R_{A,B}^2 = \sum_{i=l+1}^{d}(p_{A,i}-p_{B,i})^2$ to represent the intermediate of the left and right parts from different data custodians. Then, Eq. (5) can be represented as $d_E(p_A, p_B) = \sqrt{L_{A,B}^2 + R_{A,B}^2}$. As $L_{A,B}^2$ and $R_{A,B}^2$ can be calculated by their data owner without estimation loss, we can first calculate the intermediate result and use PrivBV for encoding. The right part can be optimized as

$$R_{A,B}^2 = \sum_{i=l+1}^{d}(p_{A,i}-p_{B,i})^2 = \underbrace{\sum_{i=l+1}^{d}(p_{A,i}^2+p_{B,i}^2)}_{S=\text{sum part}} - \underbrace{\sum_{i=l+1}^{d}2\cdot p_{A,i}\cdot p_{B,i}}_{P=\text{product part}} \qquad (6)$$

Considering that PrivBV can be used for distance estimation, we can estimate $R_{A,B}^2$ of the above equation by estimating the distance between S and P (S and P are the sum part S and product part P in Eq. (6). Letting $\mu_{\max} = 2 \times \sum_{i=l+1}^{d} \mu^2$, $R^2_{A,B}$ can be estimated as

$$\hat{R}^2_{A,B} = \frac{\mu_{\max} \cdot C_\epsilon^2}{2s} \cdot d_H(\mathcal{B}^S, \mathcal{B}^P) - \frac{\mu_{\max} \cdot \mathrm{e}^\epsilon}{(\mathrm{e}^\epsilon - 1)^2}.$$

where $\mathcal{B}^S, \mathcal{B}^P$ are the encoded bit vectors of S and P by PrivBV. The data owner of the left part can then estimate the distance between p_A, p_B as $\hat{d}_E(p_A, p_B) = \sqrt{L^2_{A,B} + \hat{R}^2_{A,B}}$, and the aggregator can estimate the distance as $\hat{d}_E(p_A, p_B) = \sqrt{\hat{L}^2_{A,B} + \hat{R}^2_{A,B}}$ (the left part proceeds in the same manner as the right part to Formula (7)). Thus, encoding and decoding occur once and twice, respectively. Compared with the original estimating method that must encode d times, this optimized method can reduce the estimation error.

4 Clustering on Anonymized Data

With the PrivBV mechanism, data from the user's side are transformed into an anonymized space with high privacy guarantees. Currently, data analysis methods on integrated datasets are limited, as the PrivBV only preserves distance information. Motivated by the k-means clustering algorithm, we present a new clustering method that only uses distance information.

The k-means algorithm[22] is one of the most fundamental clustering methods. It aims to partition all data points into k clusters by minimizing the within-cluster sum of squares (denote u_i as the mean of points in cluster C_i):

$$\arg\min \sum_{i \in [k]} \sum_{p \in C_i} (p - u_i)^2 \tag{7}$$

In each iteration of k-means, the center u_i of cluster C_i must be calculated by the mean of points in C_i. In the anonymized space, calculating the mean value is not supported by either BV or PrivBV and simply calculating the mean of bit arrays is senseless. Instead of assigning a point to its **closest center**, we assign a point to its **closest cluster** in each iteration. Given a set of observations $p_1, p_2, ..., p_n \in \mathbb{R}$, the point-to-cluster distance between point p and cluster C is defined as

$$d_C(p, C) = \sum_{p' \in C} d_E(p, p')/|C|.$$

In the anonymized space, in each dimension, the distance between an anonymous point $p \in \{0,1\}^s$ and an anonymous cluster C can be estimated as

$$\hat{d}_C(p, C) = \frac{\mu}{2s|C|} \cdot \sum_{p' \in C} \left[C_\epsilon^2 \cdot d_H(p, p') - \frac{2s \cdot \mathrm{e}^\epsilon}{(\mathrm{e}^\epsilon - 1)^2} \right].$$

Based on $\hat{d}_C$, the clustering result is given by finding the objective $C_1, C_2, ..., C_k$ such that

$$\mathop{\arg\min}_{C_1, C_2, ..., C_k} \sum_{i \in [k]} \sum_{p \in C_i} d_C(p, C_i).$$

In the distributed environment, the process of collecting and clustering in the anonymized space is shown as Algorithm 2. First, common parameters must be negotiated between users. With these parameters, each data point is encoded into a bit array with the PrivBV mechanism. As the encoding process is distance-aware, the integrated data can be used for clustering.

As with the k-means algorithm, kCluster refines the result iteratively. There are two main steps in the kCluster algorithm. In the first step, k clusters are randomly initialized by choosing k centroids and arranging each record to its nearest centroid (Line 5). In the second step, each record's label is reset to the closest cluster in the previous iteration (Lines 6–13).

Additionally, the PrivBV mechanism can be easily implemented in current clustering algorithms that only use distance information. Taking DBSCAN as an example, the essential task is to determine the number of points within the range of E.

With the PrivBV anonymization mechanism, data custodians can release their data at a high privacy-preserving level. Moreover, data utilities for clustering are guaranteed as the distance information is preserved. The proposed clustering algorithm shows the potential capacities of data mining on anonymized data.

Algorithm 2 kCluster-LDP: clustering with PrivBV

```
Input: The number of clusters $k$, encoding parameters $\mathcal{R}, t, \epsilon$,
    data range $[L, U]$ and $n$ users with their local data.
Output: The clustering result.
1: Each User encoding and releasing: $p' = \mathrm{PrivBV}_{\mathcal{R},t,\epsilon}(p)$.
2: Releasing decoding parameter: $\mu = U - L$.
3: Aggregating data: $D = p'_1 \cup p'_2 \cup \cdots$
4: Randomly choose $k$ records as initial centroids of clusters
    $C_1, C_2, ..., C_k$.
5: Assign each record $p'$ to the $I$-th cluster by
    $I = \mathop{\arg\min}_{t \in \{1,2,...,k\}} \hat{d}_C(p', C_t)$.
6: repeat
7:     Generate clusters: $\forall i \leqslant k, C'_i = C_i, C_i = \varnothing$
8:     for $j = 1, 2, 3, ..., |D|$ do
9:         $I = \mathop{\arg\min}_{t \in \{1,2,...,k\}} \hat{d}_C(p'_j, C'_t)$.
10:        $C_I = C_I \cup p'_j$.
11:    end for
12: until $\forall i \in [k], C_i = C'_i$.
13: return Set of clusters $C_1, C_2, ..., C_k$.
```

5 Experiment

In this section, we present an empirical evaluation of the proposed schemes.

5.1 Distance estimation on numerical data

To evaluate the performance of Algorithm 1, we generate a uniformly distributed dataset within the range $[0, 25]$ and implement the distance-adjusting algorithm on the estimated distance matrix of BV. The interval parameter is set as $t = 3$. Each time, 10 000 pairs are compared and evaluated. Figure 3 shows that applying the distance-adjusting algorithm can effectively lower the distance estimation error. The error is given as the MAE.

We also evaluate the performance of applying the BV and PrivBV in privacy-preserving record linkage. The effectiveness of these approaches is evaluated via F1. As shown in Fig. 4, the performance of PrivBV in record linkage is slightly worse than that of BV. This result is acceptable, as the PrivBV mechanism can provide additional privacy guarantees.

5.2 Optimization for multi-dimensional data

In this section, we consider distance estimation over multi-dimensional data. The error of the horizontally partitioned setting is not covered because it is identical to that of the non-optimization method in our experiment. For convenience, the data dimensions from different data custodians are set equal. Each record is encoded d times by PrivBV with the non-optimization method but only twice for the optimization. The range of encoded data expands during optimization, and encoding with s random values brings extra errors. Thus the number of random values we use in optimization is identical to that of non-optimization.

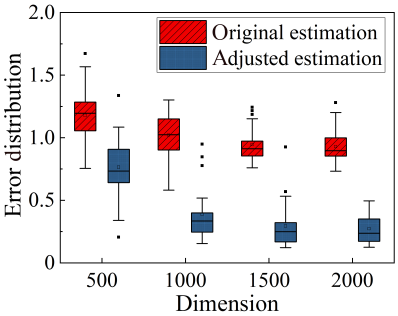


Fig. 3 Distance estimation error.

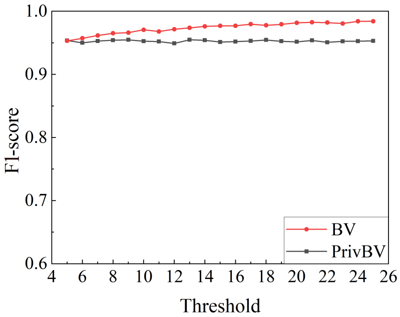


Fig. 4 PPRL performance.

Figure 5 shows that in a high dimension, the optimized method performs better. This is mainly because the non-optimized strategy must encode d times (d is the dimension of one record), whereas the optimized strategy only encodes twice. The errors accumulate in the encoding and decoding processes. Additionally, in a low dimension, the non-optimized approach performs better. This is because the interval for the random values is large in the optimized procedure. Thus the optimization strategy is ineffective under low dimension data.

5.3 Clustering performances

In this section, we evaluate the clustering performance. We use the digit dataset[23], which is composed of 1797 images, and each image is represented by an 8×8 vector. We choose k-means and kCluster as basic clustering algorithms. As far as we know, the ADP[24] and RSP[25] are the only solutions for non-interactive clustering. Unfortunately, comparing the privacy-preserving level of (ϵ, δ)-LDP, ADP, and RSP algorithms lacks a baseline. Both ADP and RSP are not theoretically privacy-preserving even though they involve noise in the encoding process. Only the PrivBV mechanism provides a rigorous privacy-preserving level theoretically.

For the ADP-based algorithm, the variance of the noise is maintained at $\sigma = 2$ and 4. For the RSP-based method, we project its dimension to 50% and 75%. The clustering result measured by NMI is listed in Table 3. It shows that the proposed mechanism can provide acceptable clustering results under strict privacy

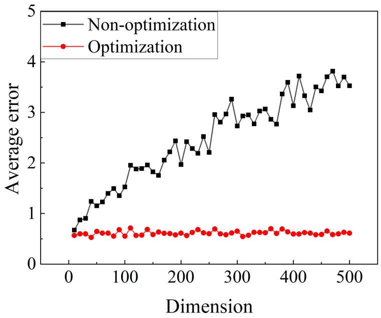


Fig. 5 Optimization analysis.

Table 3 Clustering results.

Method	Privacy level	NMI (%)
k-means	–	74.32
kCluster	–	**74.65**
RSP+k-means	50%-reduction	63.65
	75%-reduction	67.08
ADP+k-means	$\sigma = 2$	72.72
	$\sigma = 4$	62.58
PrivBV+kCluster	$(1, 8.9 \times 10^{-137})$-LDP	70.89
	$(2, 7.5 \times 10^{-56})$-LDP	**73.57**

guarantees.

In Fig. 6, we vary the epsilon ϵ in clustering and investigate its effect (with $\delta = \left(\dfrac{e^{\epsilon}}{e^{\epsilon}+1}\right)^{s} - e^{\epsilon}\cdot\left(\dfrac{1}{e^{\epsilon}+1}\right)^{s}$ and $s = 1000$). According to the experimental results, the performance fluctuates within a small range. This is because the distance information is preserved when encoding.

6 Conclusion and Discussion

This work investigates encoding mechanisms under LDP guarantees and applications in distributed scenarios. Specifically, we discover the distance-continuation property in the anonymized space and expand the BV mechanism to be locally differentially private. Our results show that we can achieve (ϵ, δ)-LDP in the anonymized space, as well as distance estimation utilities. The proposed solution can be used for application fields that depend on distance information, such as privacy-preserving data releasing and multi-party clustering in the distributed environment. As an application case, we design an algorithm for non-interactive clustering.

In this paper, we only achieve (ϵ, δ)-LDP for distance estimation and its applications. Whether ϵ-LDP can be achieved in an anonymization mechanism remains an open and challenging question.

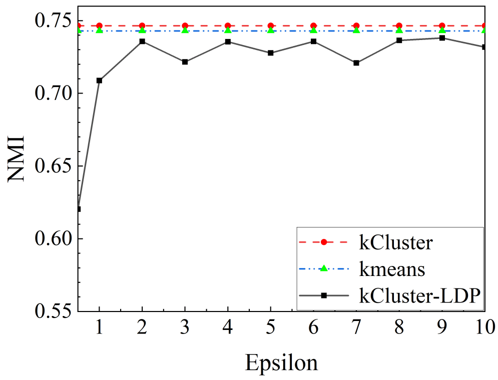


Fig. 6 Clustering results.

Acknowledgment

This work was supported by National Key Research and Development Program of China (Nos. 2019QY1402 and 2016YFB0800901).

References

[1] W. Zhang, Z. Li, and X. Chen, Quality-aware user recruitment based on federated learning in mobile crowd sensing, *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 869–877, 2021.

[2] J. Y. Hua, G. Yue, and S. Zhong, Differentially private publication of general time-serial trajectory data, in *2015 IEEE Conf. Computer Communications*, Hong Kong, China, 2015, pp. 549–557.

[3] D. Vatsalan and P. Christen, Privacy-preserving matching of similar patients, *J. Biomed. Inform.*, vol. 59, pp. 285–298, 2016.

[4] D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, FEDERAL: A framework for distance-aware privacy-preserving record linkage, *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 2, pp. 292–304, 2018.

[5] D. Karapiperis, A. Gkoulalas-Divanis, and V. S. Verykios, Distance-aware encoding of numerical values for privacy-preserving record linkage, in *2017 IEEE 33rd Int. Conf. Data Engineering*, San Diego, CA, USA, 2017, pp. 135–138.

[6] Y. Khazbak, J. Fan, S. Zhu and G. Cao, Preserving personalized location privacy in ride-hailing service, *Tsinghua Science and Technology*, vol. 25, no. 6, pp. 743–757, 2020.

[7] P. Kairouz, S. Oh, and P. Viswanath, Extremal mechanisms for local differential privacy, *J. Mach. Learning Res.*, vol. 17, no. 1, pp. 492–542, 2016.

[8] M. Fanaeepour and B. I. P. Rubinstein, Histogramming privately ever after: Differentially-private data-dependent error bound optimisation, in *2018 IEEE 34th Int. Conf. Data Engineering*, Paris, France, 2018, pp. 1204–1207.

[9] R. Bassily and A. Smith, Local, private, efficient protocols for succinct histograms, in *Proc. 47th Annu. ACM Symp. Theory of Computing*, New York, NY, USA, 2015, pp. 127–135.

[10] J. Hsu, S. Khanna, and A. Roth, Distributed private heavy hitters, in *Int. Colloquium on Automata, Languages, and Programming*, A. Czumaj, K. Mehlhorn, A. Pitts, and R. Wattenhofer, eds. Warwick, UK: Springer, 2012, pp. 461–472.

[11] G. Cormode, T. Kulkarni, and D. Srivastava, Marginal release under local differential privacy, in *Proc. 2018 Int. Conf. Management of Data*, New York, NY, USA, 2018, pp. 131–146.

[12] B. L. Ding, J. Kulkarni, and S. Yekhanin, Collecting telemetry data privately, in *Advances in Neural Information Proc. Systems*, Long Beach, CA, USA, 2017, pp. 3571–3580.

[13] C. Dwork and A. Roth, The algorithmic foundations of differential privacy, *Found. Trends Theoret. Comput. Sci.*, vol. 9, no. 3, pp. 211–407, 2014.

[14] N. Wang, X. K. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu, Collecting and Analyzing Multidimensional Data with Local Differential Privacy, in *2019 IEEE 35th Int. Conf. Data Engineering (ICDE)*, Macao, China, 2019, pp. 638–649.

[15] D. Vatsalan, P. Christen, and V. S. Verykios, A taxonomy of privacy-preserving record linkage techniques, *Informat. Syst.*, vol. 38, no. 6, pp. 946–969, 2013.

[16] T. H. Wang, J. Blocki, N. H. Li, and S. Jha, Locally differentially private protocols for frequency estimation, in *Proc. 26th USENIX Conf. Security Symp.*, Berkeley, CA, USA, 2017, pp. 729–745.

[17] C. Dwork, Differential privacy, in *Proc. 33rd Int. Conf. Automata, Languages and Programming*, Venice, Italy, 2006, pp. 1–12.

[18] B. Ding, H. Nori, P. Li, and J. Allen, Comparing population means under local differential privacy: With significance and power, in *Proc. 32nd AAAI Conf. Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 26–33.

[19] T. H. Wang, N. H. Li, and S. Jha, Locally differentially private frequent itemset mining, in *Proc. 2018 IEEE Symp. Security and Privacy*, San Francisco, CA, USA, 2018, pp. 127–143.

[20] Ú. Erlingsson, V. Pihur, and A. Korolova, RAPPOR: randomized aggregatable privacy-preserving ordinal response, in *Proc. 2014 ACM SIGSAC Conf. Computer and Communications Security*, Scottsdale, AZ, USA, 2014, pp. 1054–1067.

[21] A. Lancichinetti, S. Fortunato, and J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, *New J. Phys.*, vol. 11, no. 3, pp. 033015, 2009.

[22] J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[24] S. R. M. Oliveira and O. R. Zaiane, Privacy preserving clustering by data transformation, *J. Inf. Data Manag.*, vol. 1, no. 1, pp. 37–37, 2010.

[25] K. Liu, H. Kargupta, and J. Ryan, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 92–106, 2006.

Lin Sun received the BS degree from Huazhong University of Science and Technology, Wuhan, China in 2016. He is now a PhD candidate at the School of Software, Tsinghua University, Beijing, China. His main research interests are in the areas of privacy protection and data mining.

Guolou Ping received the master degree in computer science and technology from Hunan University, China in 2019. At present, he is a PhD candidate at the School of Software, Tsinghua University, Beijing, China. His interests include information security and network attack detection.

Xiaojun Ye received the BS degree in mechanical engineering from Northwestern Polytechnical University, China, in 1987, and the PhD degree in information engineering from INSA Lyon, France, in 1994. Currently, he is a professor at the School of Software, Tsinghua University, Beijing, China. His research interests include cloud data management, data security and privacy, and database system testing.