Received: 10 March 2019 | Revised: 30 June 2019 | Accepted: 27 August 2019
DOI: 10.1002/cpe.5607

WILEY

SPECIAL ISSUE PAPER

Heterogeneous differential privacy for vertically partitioned databases

Yang Xia[1] | Tianqing Zhu[2] | Xiaofeng Ding[1] | Hai Jin[1] | Deqing Zou[1]

[1]National Engineering Research Center for Big Data Technology and System, Service Computing Technology and System Lab, Cluster and Grid Computing Lab, Huazhong University of Science and Technology, Wuhan, China
[2]Deakin University, Victoria, Australia

Correspondence
Xiaofeng Ding, National Engineering Research Center for Big Data Technology and System, Service Computing Technology and System Lab, Cluster and Grid Computing Lab, Huazhong University of Science and Technology, Wuhan 430074, China.
Email: xfding@hust.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2016QY02D0302; National Natural Science Foundation of China (NSFC), Grant/Award Number: 61472148

Summary

Existing privacy-preserving approaches are generally designed to provide privacy guarantee for individual data in a database, which reduces the utility of the database for data analysis. In this paper, we propose a novel differential privacy mechanism to preserve the heterogeneous privacy of a vertically partitioned database based on attributes. We first present the concept of privacy label, which characterizes the privacy information of the database and is instantiated by the classification. Then, we use an information-based method to systematically explore the dependencies between all attributes and the privacy label. We finally assign privacy weights to every attribute and design a heterogeneous mechanism according to the basic Laplace mechanism. Evaluations using real datasets demonstrate that the proposed mechanism achieves a balanced privacy and utility.

KEYWORDS

differential privacy, heterogeneous privacy, privacy label, vertically partitioned data

1 | INTRODUCTION

Increasingly, data holders such as governments and industries (eg, healthcare providers), have adopted the open data initiative, where data under the custodianship of government organizations and institutions is made accessible to the public. Such shared data can then be used for statistical analysis, such as income allocation, epidemic disease prevention, and marketing strategy. However, it is important that privacy-sensitive information (eg, personally identifiable information) in such datasets are adequately protected prior to releasing the datasets.[1-3]

In the security and privacy literature, a number of privacy-preserving models and approaches (eg, *k*-anonymity, *l*-diversity, and *t*-closeness)[4-6] have been proposed. Another popular privacy-preserving model is the differential privacy (DP) proposed by Dwork et al,[7] and it guarantees that the probability of any outcome is similar between data sets that differ in one record. Specifically, DP provides a strong security guarantee of individual privacy information and has widely utilized. There are two key parameters in the DP literature, namely: sensitivity and privacy budget. The former determines how much perturbation is required in the mechanism, and the latter controls the privacy guarantee level of the mechanism. Normally, sensitivity is determined by the influence of a record on the database and privacy budget is manually set.

In this paper, we focus on the view of heterogeneous privacy-preserving based on attributes because different attributes potentially leak privacy with different risk levels, and they can influence the utility differently. For instance, a person may regard the weight as more private information than the gender, and the other person may consider gender to be more private information than the height and eye color. While protecting the privacy of the persons, we need to protect more for the weight of the former person and the gender for the latter person.

There are a number of challenges we need to address. First, how do we "label" the privacy information that we intend to preserve? Unlike *k*-anonymity, DP does not classify attributes into Quasi-Identifier (QID) or Sensitive Attributes (SA), and a privacy description is needed so that privacy is not only well protected regardless of learning method but also lays the foundation for our privacy-preserving method. The second challenge is how to find a fine-grained way to preserve privacy heterogeneously. This is important for attributes to identify their dependencies on privacy information. The third challenge is how to apply DP heterogeneously and vertically.

https://doi.org/10.1002/cpe.5607

To address these three challenges, first, we assume the concept of privacy label, which describes the characteristic of the privacy information. The privacy label can be specified according to learning methods. Second, we propose a information-based method for attributes to identify their dependencies on privacy information. We adopt the entropy to measure the information quality of privacy label and attributes. Then, we calculate the information gain ratio between them to find their relationship. Third, we design a mechanism to ensure that attributes are protected heterogeneously with DP. When given a privacy budget, we assign weights to different attributes (referred to as privacy weights) based on the information gain ratio, and we design a heterogeneous DP mechanism with those sub-privacy budgets.

From the findings of the empirical experiments using real datasets, we observe that our vertically heterogeneous DP is practical, and can achieve a better trade-off level than using conventional approaches.

In the next section, we will introduce the preliminaries used in this paper, before describing related literature in Section 6. Sections 3 and 4 describe a use case and our proposed approach, respectively. The evaluations are presented in Section 5 before this paper is concluded in Section 7.

2 | PRELIMINARIES

In this section, we will introduce some preliminaries used in this paper, including definitions of mathematics.

2.1 | Information gain ratio

Information gain ratio (IGR), ratio of information gain to the intrinsic information, is an information metric used to reduce a bias toward multi-attributes by taking the number and size of branches into account in the selection of an attribute. It can be computed using the entropy theory,[8] described as follows.

Definition 1. Given a discrete random variable X with possible values $\{x_1, \dots, x_n\}$ and probability $P(X)$, the entropy H is

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i).$$

Definition 2. Let A be the set of all k attributes and X the set of all n training examples, $v(x_i, A_j)(i = 1, \dots, n; j = 1, \dots k)$ with $x_i \in X$ defining the value of a specific example x_i for attribute $A_j \in A$, and H specifies the entropy. The $v(A_j)$ function denotes the set of all possible values of attribute $A_j \in A$. The information gain for an attribute $A_j \in A$ is defined as follows:

$$IG(X, A_j) = H(X) - H(X|A_j),$$

where $H(X|A_j)$ is the conditional entropy of X based on A_j and

$$H(X|A_j) = -\sum_{v \in v(A_j)} \left(\frac{|x_i \in X | v(x_i, A_j) = v|}{|X|} \cdot H(x_i \in X | v(x_i, A_j) = v) \right).$$

The intrinsic value for an attribute is defined as follows:

$$IV(X, A_j) = -\sum_{v \in v(A_j)} \frac{|x_i \in X | v(x_i, A_j) = v|}{|X|} \cdot \log \left(\frac{|x_i \in X | v(x_i, A_j) = v|}{|X|} \right).$$

The information gain ratio is the ratio between the information gain and the intrinsic value

$$IGR(X, A_j) = \frac{IG(X, A_j)}{IV(X, A_j)}.$$

It is generally accepted that IGR has an advantage over information gain that it is more accurate and can be used on large datasets.

2.2 | Differential privacy

Differential Privacy (DP) requires that the output of a data analysis mechanism be approximately the same, even if any single record in the input database is arbitrarily added or removed.

Definition 3 (ϵ-Differential Privacy[7]). A randomized mechanism $\mathcal{A}$ gives ϵ-differential privacy if for any pair of neighboring datasets D_1 and D_2, and any $S \subseteq Range(\mathcal{A})$,

$$Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{A}(D_2) \in S].$$

This protects the privacy of any single record because adding or removing any single record results in e^{ϵ}-multiplicative-bounded changes in the probability distribution of the output.

Achieving differential privacy revolves around hiding the presence or absence of a single individual. Consider the query "How many rows in the database satisfy property P?" The presence or absence of a single row can affect the answer by at most 1. Thus, a differentially private mechanism for a query of this type can be designed by first computing the true answer and then adding random noise according to a distribution with the following property:

$$\forall z, z' s.t|z - z'| = 1 : Pr[z] \leq e^{\epsilon} Pr[z'].$$

Definition 4 (Sensitivity[9]). For $f : \mathcal{D} \rightarrow \mathbf{R}^d$, the L_1 sensitivity of f is

$$\Delta f = \max_{D_1,D_2} ||f(D_1) - f(D_2)||_1$$
$$= \max_{D_1,D_2} \sum_{i=1}^{d} ||f(D_1)_i - f(D_2)_i||,$$

for all D_1, D_2 differing in at most one row.

Definition 5 (Laplace mechanism[7]). The Laplace distribution with parameter b, denoted $Lap(b)$, has the density function

$$P(z|b) = \frac{1}{2b} exp(-|z|/b).$$

Its variance is $2b^2$.

Taking $b = 1/\varepsilon$, the density at z is proportional to $e^{-\varepsilon|z|}$. For any z, z' such that $|z - z'| \leq 1$, the density at z is at most e^{ε} times the density at z', satisfying the condition in Equation 1.

Theorem 1. *For $f : \mathcal{D} \rightarrow \mathbf{R}^d$, the mechanism $\mathcal{M}$ that adds independently generated noise with the distribution $Lap(\Delta f/\varepsilon)$ to each of the d output terms enjoys ε-differential privacy.*[9]

2.3 | Trade-off metric

Our work provides privacy protection from the view that different attributes contain different privacy information, mainly based on the notion that the utility of the data should not be significantly reduced because of privacy protection. We use the trade-off metric proposed by Fung et al[10] to quantify our approach and optimize the method of privacy-preserving.

Definition 6. Given a generation g, the optimization is to minimize

$$ILPG(g) = \frac{IL(g)}{PG(g) + 1},$$

where $IL(g)$ denotes the information loss and $PG(g)$ denotes the privacy gain by performing g. The choice of $IL(g)$ and $PG(g)$ depends on the information metric and privacy model.

3 | CASE STUDY AND PROBLEM DEFINITION

In this section, we study the case to demonstrate the need to heterogeneously preserve privacy of attributes and describe the problem definition of our work.

3.1 | Case study

For understandability, we use k-anonymity to study the case, and it is also appropriate for DP. In the most basic form of anonymization, private data can be divided into Quasi-Identifier (QID) and Sensitive Attributes (SAs). QID is a set of attributes that could potentially identify record owners, and SA consists of sensitive person-specific information such as disease, salary, and disability status.

Suppose there is a private data table (see Figure 1A) and an external data table (see Figure 1B), and in the private data, the set of QID is {Job, Sex, Age} and the set of SA is {Disease}.

It is trivial to note that each record cannot be distinguished under only a single QID attribute. The private data leaks privacy when it is linked with the external data only under two or three QID attributes. In other words, we are not able to accurately infer an individual through the age of neither 30, 35, nor 38. However, we can accurately infer an individual by combining job (ie, dancer), sex (ie, female), and age (ie, 30).

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(A)

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	35
Irene	Dancer	Female	32

(B)

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(C)

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

(D)

FIGURE 1 Example illustrating *k*-anonymity. A, Private data; B, External data; C, 3-anonymous private data; D, 4-anonymous external data

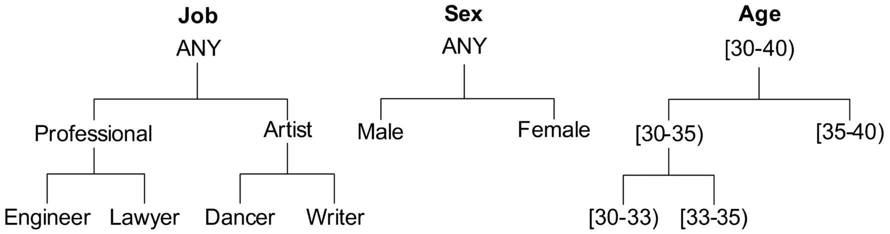


FIGURE 2 Taxonomy trees for QID

Thus, we should consider privacy-preserving of all attributes. Based on a single attribute, we can increase the anonymous degree scale. The example generalizes private and external data to satisfy 3-anonymous (see Figure 1C) and 4-anonymous (see Figure 1D), respectively, with Taxonomy trees in Figure 2.

All attributes of *k*-anonymity are protected at the same level. Particularly, each record of attributes in *QID′* that is anonymous are all at least indistinguishable from k-1 same records. However, we consider that different attributes should not be protected at the same level for two reasons.

1. Individuals are distinguished with different maximum probabilities under every attribute. This means the risk of privacy leakage for every vertical attribute is different. In our three attributes example, an individual is identified with probability at most 100% (Job), 17% (sex), and 50% (Age), respectively.
2. Different attributes may influence sensitive information of individuals differently. In other words, given a sensitive attribute, the conditional probabilities that individuals of the other attributes are different (see Figure 3). This is important as it could influence the utility of private data.

Definition 7. Let $|QID|_1, |QID|_2$ denote the corresponding count of *QID* value of private data and external data, respectively. $|SA|$ denotes the count of *SA* value, and $|QID_1 \cap SA|$ denotes the count of *QID* value joined with *SA* value in the private data. We define the maximum identifiable probability of *QID* to be

$$mPr[QID] = \max \frac{|QID_1 \cap SA|}{|QID_2|}.$$

The conditional probability of *QID* based on *SA* is

$$cPr[QID|S] = \frac{|QID_1, SA|}{|SA|}.$$

QID \ SA		Disease		
		Hepatitis	HIV	Flu
Job	Engineer	100%	0%	0%
	Lawyer	0%	25%	0%
	Writer	0%	25%	100%
	Dancer	0%	50%	0%
Sex	Male	100%	25%	0%
	Female	0%	75%	100%
Age	30	0%	75%	100%
	35	50%	0%	0%
	38	50%	25%	0%

FIGURE 3 Conditional probabilities of *QID* based on *SA*

The aforementioned case study illustrates the need for vertical heterogeneity in privacy-preserving via k-anonymous, as well as DP (since sensitivity and privacy budget are two critical parameters in DP-based approaches). The sensitivity determines how much perturbation is required in the privacy-preserving mechanism and privacy budget controls the privacy guarantee level of the mechanism. Similarly, for all attributes, their sensitivities may differ and privacy budgets also need to be selected differently.

3.2 | Problem definition

In this section, we describe the attack model in our scene and illustrate the problem to solve by definition.

The attack model is simple in our scene. Actually, the data owner curator provides data to the data analyst or consumer for a certain task. The data privacy can be violated by data analyst or consumer during the learning process. We do not care what method does the data analyst or consumer use but only need to fix the learning task. In this way, the data analyst or consumer is semi-honest since the learning task should not be impossible to be finished.

Suppose there is a dataset D; we partition it vertically to j different sub-datasets $D_1, D_2, \dots, D_j$ according to attributes. Each sub-dataset contains the private information of the individuals, and the private information may be less or more sensitive compared among different sub-datasets. Our goal is to preserve the individuals' privacy of sub-datasets with different levels while the dataset D does not leak privacy.

4 | PROPOSED APPROACH

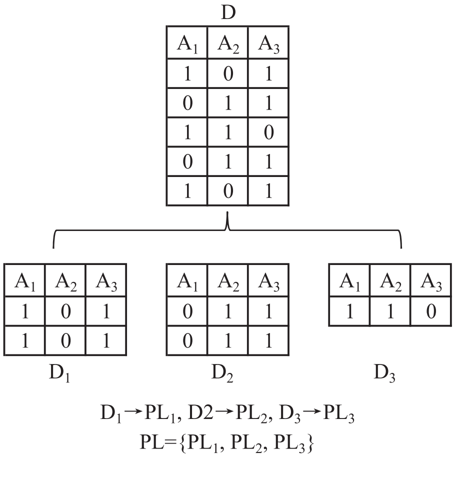
In our proposed approach, we first introduce the concept of privacy label, which characterizes the privacy information of the learning results. Then, we develop a information-based framework to systematically explore the dependencies between all attributes and the privacy label inspired by Sadhya et al,[11] and select the attribute data of the different dependencies to design a heterogeneous differential privacy mechanism, hereafter referred to as vertically heterogeneous differential privacy (VHDP).

We will describe our approach in the remaining of this section.

4.1 | Characterize privacy

In privacy-preserving data publishing, the data publisher has a table denoted by explicit identifier, quasi identifier, sensitive attributes, and non-sensitive attributes.[12] Sensitive attributes are usually protected as private information (ie, in most approaches, sensitive attributes are regarded as private information).

However, when a dataset does not have an obvious privacy attribute, such private information "hidden" in the dataset could potentially be exposed via data learning. For example, an adversary wishes to learn more information about the multi-attributes database. Using classification, individual observations are analyzed into a set of quantifiable properties (ie, explanatory variables or features), which can be categorical (eg, A, B, AB, or O for blood type), ordinal (eg, large, medium, or small), integer-value (eg, number of occurrences of a partial word in an email), real-value (eg, measurement of blood pressure), and/or Boolean value (eg, true or false, yes or no). Generally, these properties are known as classification labels. For different attributes corresponding to classification labels, the risk of privacy breach may differ as an adversary's capability to target privacy information with each attribute may vary. If the learning method is known, then we classify the learning results according to the leaning method and regard all distinct categories as privacy labels. If the learning method is unknown without learning results, then we regard the privacy labels as null. The definition of the privacy label for a known learning method is defined as follows and an example is given in Figure 4.



D

A_1	A_2	A_3
1	0	1
0	1	1
1	1	0
0	1	1
1	0	1

D_1

A_1	A_2	A_3
1	0	1
1	0	1

D_2

A_1	A_2	A_3
0	1	1
0	1	1

D_3

A_1	A_2	A_3
1	1	0

$D_1 \rightarrow PL_1$, D2 $\rightarrow PL_2$, $D_3 \rightarrow PL_3$

$PL=\{PL_1, PL_2, PL_3\}$

FIGURE 4 Example of privacy label

Definition 8 (Privacy label). Given a dataset $D = \{A_1, A_2, \ldots, A_k\}$ with a method $\mathcal{M}$, the dataset contains k attributes and is classified into m categories $D_j (j = 1, \ldots, m)$ according to $\mathcal{M}$, $D = \cup_{j=1}^{m} D_m$, and $D_p \cap D_q = \phi (p \neq q)$. The privacy label set is $PL = \{PL_j | \mathcal{M} : D_j \to PL_j\} (j = 1, \ldots, m)$.

Our work mainly focuses on privacy-preserving instead of learning methods and our approach is general, so we consider that PL is not null.

4.2 | Calculate information gain ratio

We calculate the information gain ratio (IGR) scores between different attributes and PL to evaluate their influence on privacy. For simplicity, we suppose that the dataset is a discrete uniform distribution, where the probability value can be calculated based on the count.

Definition 9. Given a data set D of n data records that has been labeled with a privacy label set PL, PL contains m labels $PL = (PL_1, PL_2, \ldots, PL_m)$, and every label $PL_i (i = 1, 2, \ldots, m)$ is counted as $pl_i = |PL_i|$ and the total count of privacy labels is $|PL| = n$. Then, the entropy of privacy labels is

$$H(PL) = \begin{cases} -\sum_{i=1}^{m} \frac{pl_i}{n} \cdot \log \frac{pl_i}{n} & PL \neq \emptyset \\ 0 & PL = \emptyset. \end{cases}$$

Definition 10. Given a data set D of n data records, its attribute set A contains k attributes $A = (A_1, A_2, \ldots, A_k)$. For every attribute A_i of D, we count the number set $T_i = (t_{i1}, t_{i2}, \ldots, t_{ij}) (i = 1, 2, \ldots, k)$ of j different attribute values $V_i = (v_{i1}, v_{i2}, \ldots, v_{ij})$, ie, the ith attribute has j different values, and every value has $t_{i1}, t_{i2}, \ldots t_{ij}$ data records, respectively. Then, the intrinsic value for an attribute i is

$$IV(PL, A_i) = -\sum_{s=1}^{j} \frac{t_{is}}{n} \cdot \log \frac{t_{is}}{n}.$$

For instance, if A_1 has values of $(1, 0, 0, 1, 0)$, then we obtain $j = 2$, $v_{11} = 1$, $v_{12} = 0$, $V_1 = (1, 0)$, $t_{11} = 2$, $t_{12} = 3$, and $T_1 = (2, 3)$.

Definition 11. Given j different attribute values V_i of A_i, a function $F : V_i \times PL \to R^{j \times m}$ is

$$F(V_i, PL) = \begin{pmatrix} t'_{11} & t'_{12} & \cdots & t'_{1m} \\ t'_{21} & t'_{22} & \cdots & t'_{2m} \\ \cdots & \cdots & \cdots & \ldots \\ t'_{j1} & t'_{j2} & \cdots & t'_{jm}. \end{pmatrix}$$

The element $t'_{sr} (s = 1, 2, \ldots, j; r = 1, 2, \ldots, m)$ in the aforementioned matrix denotes the number of every attribute value v_{is} that is contained in every privacy label PL_r. Then, the information gain of attribute i is as follows:

$$IG(PL, A_i) = H(PL) - H(PL|A_i),$$

where the conditional entropy of the privacy label based on an attribute is

$$H(PL|A_i) = -\sum_{s=1}^{j} \frac{t'_{is}}{n} \sum_{r=1}^{m} \frac{t'_{sr}}{t'_{is}} \log \frac{t'_{sr}}{t'_{is}}$$

and IGR of attribute i is

$$IGR(PL, A_i) = \frac{IG(PL, A_i)}{IV(PL, A_i)}.$$

IGR scores for all attributes can be computed using Algorithm 1.

Algorithm 1 IGR score calculation for attributes of the dataset

```
$D = (A_1, A_2, \dots, A_k)$, records number of dataset $n = len(D)$;
privacy label set $PL = (PL_1, PL_2, \dots, PL_m)$;
$PL_1 \leftarrow pl_1, PL_2 \leftarrow pl_2, \dots, PL_m \leftarrow pl_m$;
the entropy of privacy label
    $H(PL) = -\sum_{i=1}^{m} \frac{pl_i}{n} \, log_2 \frac{pl_i}{n}$;
for every attribute $A_i$ do
    get the set of attribute value $V_i = (v_{i1}, v_{i2}, \dots, v_{ij})$;
    count the number set of different attribute value
        $T_i = (t_{i1}, t_{i2}, \dots, t_{ij})$;
    $V_i \leftarrow T_i$;
    intrinsic value for an attribute
        $IV(PL, A_i) = -\sum_{s=1}^{j} \frac{t_{is}}{n} log_2 \frac{t_{is}}{n}$;
    $F(V_i, PL) \rightarrow (t'_{sr})_{j \times m}$;
    the conditional entropy of privacy label on an attribute
        $H(PL|A_i) = -\sum_{s=1}^{j} \frac{t'_{is}}{n} \sum_{r=1}^{m} \frac{t'_{sr}}{t'_{is}} \, log \, \frac{t'_{sr}}{t'_{is}}$;
    information gain of an attribute
        $IG(PL, A_i) = H(PL) - H(PL|A_i)$;
    information gain ratio of an attribute
        $IGR(PL, A_i) = IG(PL, A_i)/IV(PL, A_i)$;
end for
$IGR(PL, A) = (IGR(PL, A_1), \dots, IGR(PL, A_k))$;
```

We will use the *IGR* scores in the design of our vertically heterogeneous DP mechanism.

4.3 | Vertically heterogeneous DP mechanism

As previously discussed, in this paper, we only consider the case that the privacy label is not null. In the work of Alaggan et al,[13] a privacy weight vector was used to satisfy the different privacy-preserving demands of users. However, the work limits the value of privacy to 0 and 1. In other words, it must predefine a maximum privacy budget and all other users should have a privacy budget less than this predefined budget. In this paper, we adopt a similar privacy weight vector selected method as follows.

Definition 12. Given *IGR* scores vector $\overrightarrow{IGR(PL, A)}$ of *PL* based on all attributes *A*, a function $\mathcal{F} : \overrightarrow{IGR(PL, A)} \rightarrow \vec{w}, \vec{w} = (w_1, w_2, \dots, w_k)$ denotes that the *j*th attribute has a privacy weight w_j sized by its *IGR* score, and $w_j \in [0, 1](j = 1, 2, \dots, k)$.

For example, $\vec{w} = ((\frac{1}{2})^p, \dots, \frac{1}{2}, 1, 2, \dots, 2^q)), p + q + 1 = k, 0 \leq q, p$. In other words, if an attribute is assigned privacy weight of 1, then the privacy weight of an attribute that has a lower *IGR* score is assigned less than 1 in the order of *IGR* scores and vice versa. Therefore, if the privacy budget is ϵ, then we specify the sub-privacy budget ϵ_i of an attribute A_i where its privacy weight is 1. We can see that the privacy weight vector $\vec{w}$ is selected differently in *k* ways. Every selection refers to one attribute with privacy weight 1.

When we assign multi-attributes data into several portions with a sub-privacy budget vector $\vec{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_k)$ being used for all *k* attributes, the entire dataset satisfies the bucket effect as follows.

Theorem 2 (Vertically Heterogeneous Parallel Composition). *Let $\mathcal{A}_1$ be a ϵ_1-differentially private mechanism and $\mathcal{A}_2$ be a ϵ_2-differentially private mechanism ($\epsilon_1 \neq \epsilon_2$). Let (D_1, D_2) be a vertical partition of the database D and the two disjoint databases have the same sensitivity. Then, the vertical parallel composition mechanism of D*

$$\mathcal{A}(D) = (\mathcal{A}_1(D_1), \mathcal{A}_2(D_2))$$

satisfies ϵ-differentially private, where $\epsilon = \max(\epsilon_1, \epsilon_2)$.

Proof. Given $a > 0, b > 0, c > 0, d > 0$ and $\frac{a}{b} < \frac{c}{d}$, we can easily get that $\frac{a+c}{b+d} < \frac{c}{d}$. Therefore, according to the definition of DP, there is $\frac{Pr[(\mathcal{A}_1(D_1)+\mathcal{A}_2(D_2))\in S]}{Pr[(\mathcal{A}_1(D'_1)+\mathcal{A}_2(D'_2))\in S]} = \frac{Pr[\mathcal{A}_1(D_1)\in S]+Pr[\mathcal{A}_2(D_2)\in S]}{Pr[\mathcal{A}_1(D'_1)\in S]+Pr[\mathcal{A}_2(D'_2)\in S]} \leq \max(\frac{Pr[\mathcal{A}_1(D_1)\in S]}{Pr[\mathcal{A}_1(D'_1)\in S]}, \frac{Pr[\mathcal{A}_2(D_2)\in S]}{Pr[\mathcal{A}_2(D'_2)\in S]}) \leq \max(e^{\epsilon_1}, e^{\epsilon_2})$.

Due to the vertical parallel composition, our solution can be derived to satisfy ϵ-differential privacy of *D* as a whole, where $\epsilon = \max(\epsilon_1, \epsilon_2, \dots, \epsilon_k)$. Thus, for heterogeneous DP in multi-attributes data, it has the bucket effect that the privacy-preserving level of whole data is the weakest among all attributes. Thus, the value domain of $\vec{w}$ is limited in $[0, 1]^k$.

Additionally, in the work of Alaggan et al,[13] the privacy weight is assigned horizontally. If the weight vector is disclosed, then the adversary easily knows the privacy demands of different individual records, which compromises privacy. Thus, it is necessary to hide the privacy weight learned by their learning method. In our vertical attribute data approach, we do not need to hide the weight vector because the adversary only knows the level at which an attribute contributes to the privacy label, but not the corresponding individual record.

After assigning different privacy weights to different attributes, we add noise to them one by one with DP. In our work, we adopt the Laplace mechanism to realize DP. The Laplace mechanism $\mathcal{A}$, $\mathcal{A}(D) = f(D) + Lap(b)$ with mean 0 and standard deviation b, provides

$$Pr[\mathcal{A}(D) = t] \leq e^{\epsilon} \cdot Pr[\mathcal{A}(D') = t],$$

where $b = \frac{\Delta f}{\epsilon}$. DP can be achieved by setting the perturbation to be proportional to the sensitivity and privacy budget ϵ. In this paper, we set the function f as count query function and, thus, the sensitivity is $\Delta f = 1$ for each attribute, and parameter $b = \frac{1}{\epsilon}$. □

According to the Laplace mechanism, the noise added to an attribute A_i is as follows.

Definition 13. Let Y be the matrix of noise; $Y = (y_{ij})_{n\times 1}$, y_{ij} is the noise added to every attribute value. $U = (u_{ij})_{n\times 1}$ is a random matrix that $U = rand(0, 1) - 0.5$, where $rand(0, 1)$ is a random matrix whose elements are randomly produced between 0 and 1. The mean of the Laplace distribution is $\mu = 0$ and the standard deviation is b. Then,

$$Y = \mu - b \cdot sgn(U)log(1 - 2|U|).$$

For different attributes, we assign different privacy weights in $\vec{w}$, and given a privacy budget ϵ, the sub-privacy budget vector $\vec{\epsilon} = \epsilon\vec{w}$. From this, we obtain the parameter vector of b, $\vec{b} = \frac{1}{\vec{\epsilon}}$. The heterogeneous perturbation for dataset vertically is described in Algorithm 2.

Algorithm 2 Vertically heterogeneous DP with Laplace mechanism

```
$D = (A_1, A_2, \dots, A_k)$;
sub-privacy budget vector $\vec{\epsilon} = \epsilon\vec{w} = (\epsilon_1, \epsilon_2, \dots, \epsilon_k)$;
$A_1 \leftarrow \epsilon_1, A_2 \leftarrow \epsilon_2, \dots, A_k \leftarrow \epsilon_k$;
mean of Laplace distribution $\mu = 0$;
records number of dataset $n = len(D)$;
noise matrix of an attribute $Y = (y_{ij})_{n\times 1}$;
a random matrix $U = (u_{ij})_{n\times 1}$;
for every attribute $A_i$ do
    $b_i = 1/\epsilon_i$;
    $U = rand(1, n) - 0.5$;
    if $u_{ij} > 0$ then
        $y_{ij} = \mu - b_i \cdot log(1 - 2u_{ij})$;
    else
        $y_{ij} = \mu + b_i \cdot log(1 + 2u_{ij})$;
    end if
    noisy attribute $A'_i = A_i + Y$;
end for
$D' = (A'_1, A'_2, \dots, A'_k)$;
```

Generally, when we assign privacy weights to different attributes, *IGR* scores may be very similar or even the same among some attributes. Thus, we classify *IGR* scores into different intervals and the elements in sub-privacy budget vector may be the same. If all sub-privacy budgets are the same, then our solution transforms to homogeneous DP.

4.4 | Trade-off measurement

We attempt to quantify the trade-off level of our VHDP mechanism using the trade-off metric ILPG. To calculate ILPG, the choices of $IL(g)$ and $PG(g)$ depend on the information metric and privacy model. In our work, information metric is information gain and the privacy model is DP. Thus, $IL(g)$ is the difference between the information gain of the privacy label for attributes before and after the generation and can be replaced instead by the IGR score. In our evaluations, we also validate this to determine whether it is feasible as we expect. $PG(g)$ is the differential privacy

TABLE 1 Datasets

Datasets	Number of Instances	Number of Attributes
Statlog	270	13
Adult	30,162	14

level which can be replaced instead by the reciprocal of the privacy budget ϵ (ie, $\frac{1}{\epsilon}$). The particular definition of ILPG for an attribute A_i is as follows and Algorithm 3 is used to compute ILPG for all attributes.

Algorithm 3 Measurement of ILPG scores

```
$D = (A_1, A_2, \dots, A_k), D' = (A'_1, A'_2, \dots, A'_k)$;
$\vec{\epsilon} = (\epsilon_1, \dots, \epsilon_k)$;
calculate the $IGR$ scores of attributes in $D, D'$ with Algorithm 1:
      $IGR(PL, A) = (IGR(PL, A_1), \dots, IGR(PL, A_k)), IGR(PL, A') = (IGR(PL, A'_1), \dots, IGR(PL, A'_k))$;
for every attribute $A'_i$ do
    $ILPG(PL, A'_i) = \frac{IGR(PL, A'_i) - IGR(PL, A_i)}{1 + \frac{1}{\epsilon_i}}$;
end for
$ILPG(PL, A') = (ILPG(PL, A'_1), \dots, ILPG(PL, A'_k))$;
```

Definition 14. Let $D' = (A'_1, \dots, A'_k)$ be the noisy dataset of $D = (A_1, \dots, A_k)$. $IGR(PL, A_i), IGR(PL, A'_i)$ are the IGR scores of the attribute A_i, A'_i, respectively, based on privacy label PL, and the sub-privacy budget vector $\vec{\epsilon} = (\epsilon_1, \dots, \epsilon_k)$. Then, the ILPG score of A'_i is

$$ILPG\left(A'_i\right) = \frac{IGR\left(PL, A'_i\right) - IGR(PL, A_i)}{1 + \frac{1}{\epsilon_i}}.$$

4.5 | Discussion

In our approach, we consider the separation protecting of dataset vertically using DP (ie, our proposed VHDP). VHDP adds fine-grained noise, which not only satisfies the heterogeneous privacy preservation of attributes, but also keeps non-private information data.

Sensitivity. Sensitivity is related to the function f, and the sensitivity of VHDP must satisfy not only multi-local sensitivities but also global sensitivity. Our VHDP supposes all local sensitivities and global sensitivity to be 1. This limits the function f to the type of function, such as a count query function like "How many records satisfy property P?" or "How many records satisfy property P and/or Q?" In other words, the maximum difference between the output results of the same function acting on two adjacent sets (differing in at most one row) should be 1.

Time Complexity. From the algorithms presented in this section, the complexity of our method is only increased by the longitudinal attribute scoring, and it is easy to observe that the time complexity of Algorithm 1 is $O(kn^2)$; thus, the time complexity of Algorithm 1 is concerned with the horizontal and vertical scale of the dataset. In Algorithm 2, when noise is added to the dataset, as each attribute value must have noise added one by one, the time complexity is always the same and is $O(kn)$. The time complexity of Algorithm 3 is $O(k)$. All three algorithms are polynomial time.

5 | EXPERIMENT SETUP AND FINDINGS

We use two multi-attributes datasets of different sizes to evaluate our approach, namely, the Statlog heart disease dataset (available on http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29), and the dataset of Adult (available on http://archive.ics.uci.edu/ml/datasets/Adult); see also Table 1. Both datasets are learned by classification and we regard the learning results as private information (ie, privacy label is not null).

According to our information-based evaluation approach, we first obtain the IGR scores of all attributes in the datasets. Then, we assign different sub-privacy budgets $\vec{\epsilon} = \epsilon\vec{w}$ with privacy weight $\vec{w} = ((\frac{1}{2})^p, \dots, \frac{1}{2}, 1), p + 1 = k$ and $0 \leq p$, to different attribute data according to the scores from high to low. The larger the IGR score of an attribute, the smaller the assigned privacy weight. In fact, the scores may be very close to each other in the attributes. To solve this problem, we divide them into different groups instead. After adding randomized Laplace noise to the vertical dataset to satisfy heterogeneous DP, we calculate the IGR scores of the noisy dataset. We finally calculate the ILPG scores of the noisy dataset with VHDP as well as homogeneous DP.

5.1 | Results of IGR

5.1.1 | Assign privacy weight

The IGR scores of different attributes of the two datasets are, respectively, displayed in Figure 5A and Figure 5B. As we observe, there are 13 IGR scores for Statlog and 14 IGR scores for Adult, corresponding to different attributes ID. In both figures, every attribute ID denotes the attribute we manually marked for each dataset.

We observe that the IGR scores are negative numbers. This is because the entropy of privacy labels is less than the entropy of privacy labels based on every attribute. In other words, a single attribute has more information than the privacy labels, which are classified among the learning results of all attributes. The larger the IGR score of an attribute, the greater its dependency on privacy labels; thus, we need to ensure stronger privacy protection for it.

For both Statlog and Adult datasets, we divide their attributes into three groups with three IGR score intervals to assign different privacy weights. The privacy weight vector is $\vec{w} = ((\frac{1}{2})^p, (\frac{1}{2})^{p-1}, (\frac{1}{2})^{p-2})$, $p = 2$. In this case, the parameter p is selected from group number minus 1, so we have only one heterogeneous way for every dataset. In other words, for different privacy budgets, the attributes have immutable privacy weights. For example, given neither a stable privacy budget $\epsilon = 0.2$ or $\epsilon = 0.3$, the heterogeneous way is assigned with the weight $\vec{w} = (\frac{1}{4}, \frac{1}{2}, 1)$.

5.1.2 | Validation and utility

With different privacy weights, we use the Laplace mechanism to perturb the real data and achieve DP. In order to validate the relationship between IGR score and the amount of noise or utility, we compare the IGR scores of the real data and the noisy data perturbed by VHDP as well as homogeneous DP with different privacy budgets. All results are displayed in Figures 6 and 7.

In both Figures 6 and 7, epsilon ϵ denotes the privacy budget (provided in advance), and we use four privacy budgets {0.1, 0.2, 0.3, 0.4}. For every privacy budget, the label N/A denotes the IGR score of real data without adding noise, which is invariant. VHDP and DP denote the IGR scores of the noisy datasets in different privacy weight vectors. VHDP denotes the IGR score of the VHDP with the privacy weight vector

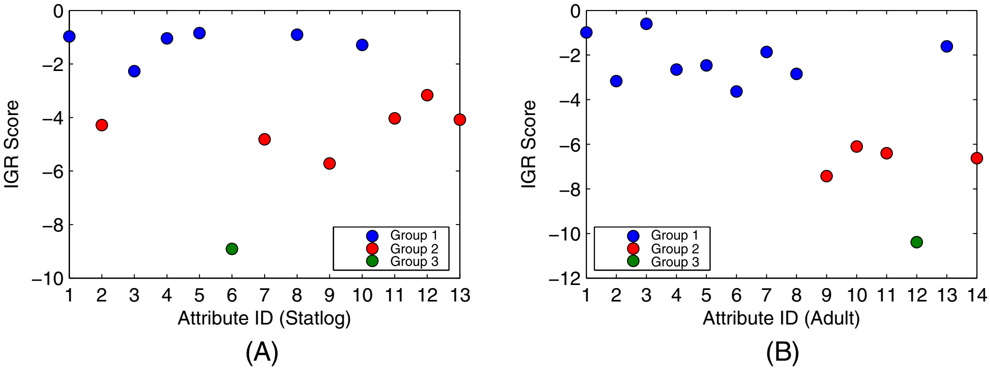


FIGURE 5 Findings of IGR evaluation

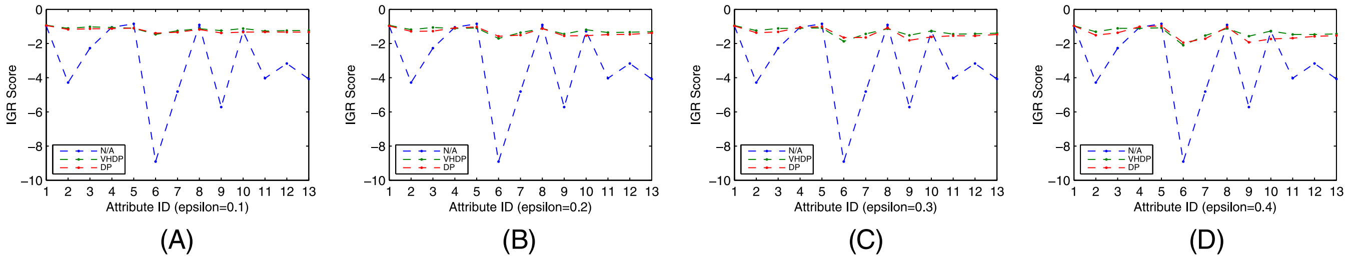


FIGURE 6 Findings of IGR scores under different privacy budgets (Statlog)

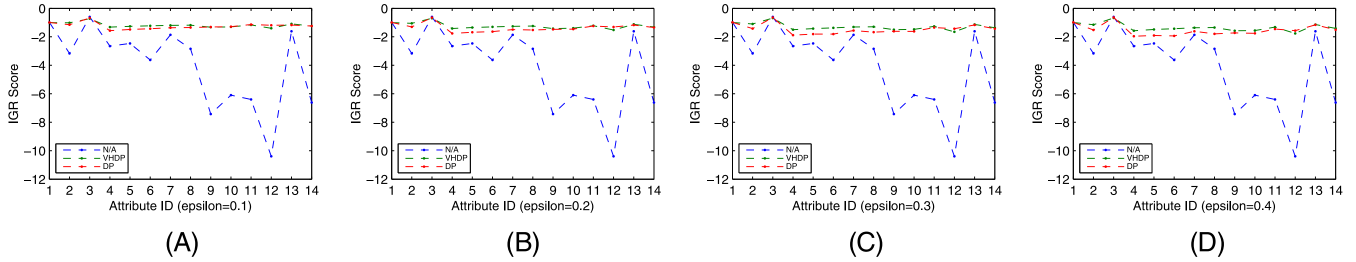


FIGURE 7 Findings of IGR scores under different privacy budgets (Adult)

$\vec{w} = (\frac{1}{4}, \frac{1}{2}, 1)$, and DP denotes the IGR score of homogeneous DP with the privacy weight vector $\vec{w} = (1, 1, 1)$. As observed, all IGR scores of noisy datasets are also negative numbers as N/A, and the IGR scores of N/A for all attributes are significant less than those of VHDP and DP.

The findings show that both the heterogeneous and homogeneous approaches increase the IGR score compared with N/A. This is because the information uncertainty of privacy labels based on attributes is reduced after adding noise (ie, the attribute data has less information; thus, reducing the data's utility). This supports our argument that the utility of data is weakened after adding noise. Thus, we can conclude that the IGR score is positively correlated with information loss. This also indicates that we can use IGR scores of attributes to express information loss to measure the trade-off level with the ILPG metric.

Utility. In order to test the utility of attributes, we use bias error of IGR scores before and after perturbation. The results are shown in Figures 8 and 9 for different privacy budgets. The findings show that VHDP does not decrease the utility of attributes much especially compared with DP.

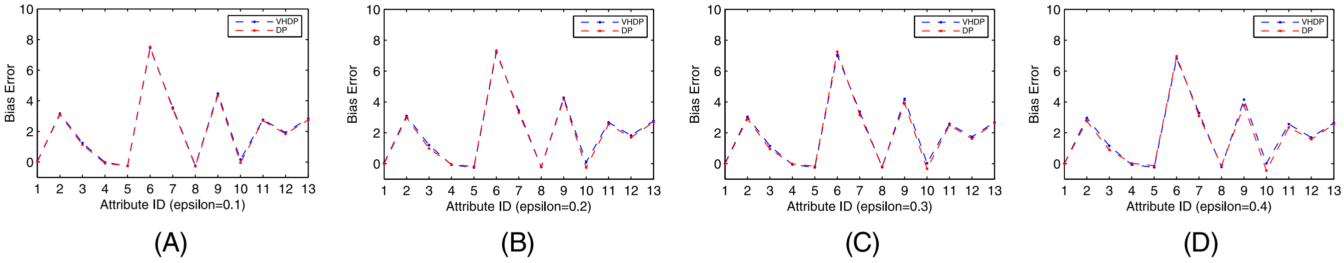
5.2 | Results of ILPG

It is easy to determine that the ILPG score of all attributes is zero when no noise is added. If the dataset is perturbed, then the ILPG score must be greater than zero.

5.2.1 | VHDP

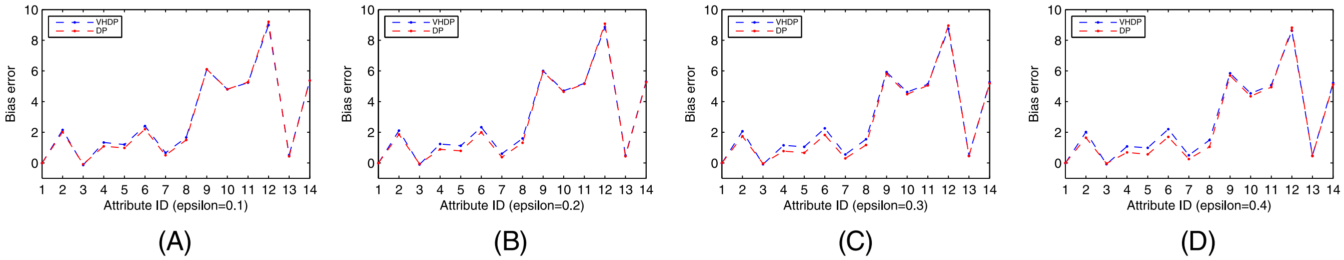
Using our proposed method, both Statlog and Adult datasets are perturbed in the way of VHDP as discussed above. We use the metric of ILPG to measure it and find the relationship of our vertically heterogeneous DP and the privacy budget. The results are displayed in Figure 10A for Statlog and Figure 10B for Adult.

In both Figures 10A and 10B, we display the ILPG scores of different attributes with different privacy budgets {0.1, 0.2, 0.3, 0.4}. The four labels represent different privacy budgets, and every colored mark represents the ILPG score of one attribute. As observed, the privacy budget is smaller, and the ILPG scores of different attributes are lower. Clearly, the ILPG score is at a minimum with the least privacy budget ϵ=0.1 for every attribute.



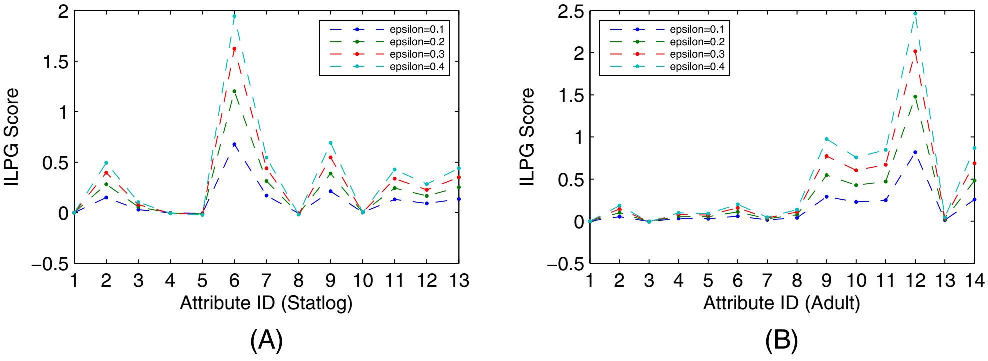
(A) (B) (C) (D)

FIGURE 8 Findings of bias error under different privacy budgets (Statlog)



(A) (B) (C) (D)

FIGURE 9 Findings of bias error under different privacy budgets (Adult)



(A) (B)

FIGURE 10 Findings of ILPG scores under different privacy budgets using VHDP

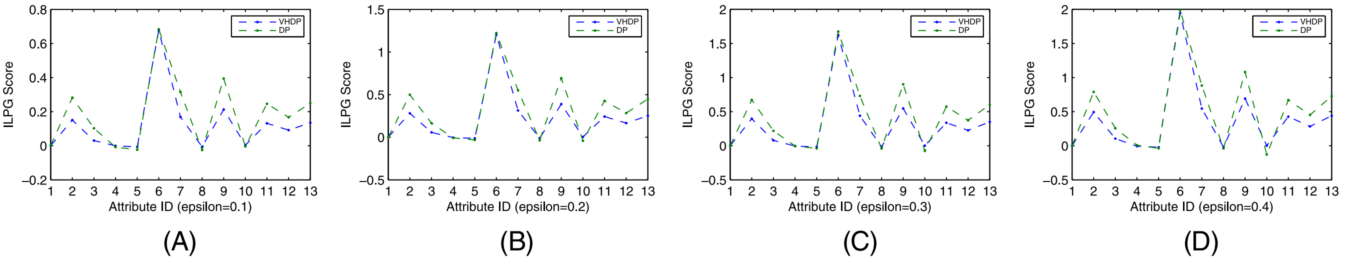


FIGURE 11 Findings of ILPG scores under different privacy budgets using VHDP and DP (Statlog)

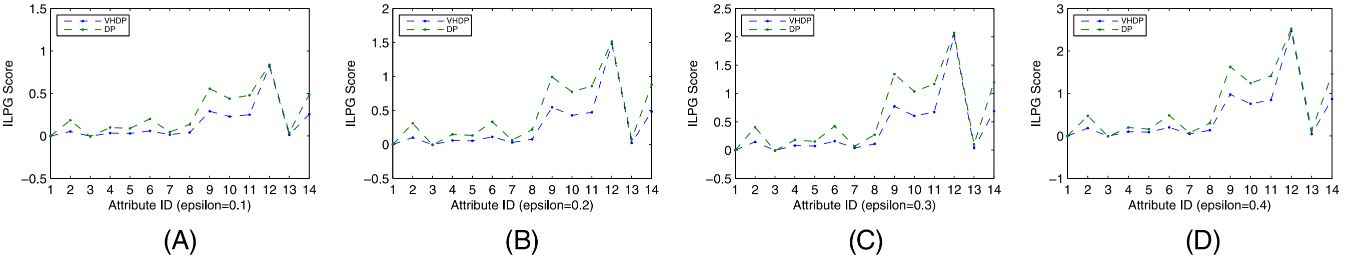


FIGURE 12 Findings of ILPG scores under different privacy budgets using VHDP and DP (Adult)

The results show that, when the privacy budget is smaller, the vertically heterogeneous DP achieves a better trade-off level. As we know, the privacy budget of DP represents the privacy-preserving level of the dataset. A smaller privacy budget means that the DP provides stronger privacy preservation. Thus, our method achieves the goal of a better trade-off level as well as stronger privacy preservation.

5.2.2 | VHDP versus DP

To compare the trade-off level of our VHDP with conventional homogeneous DP, we use ILPG to measure VHDP and conventional homogeneous DP. The results are displayed in Figure 11 for Statlog and Figure 12 for Adult. In both figures, we display the value of the ILPG scores of our vertically heterogeneous DP as well as homogeneous DP with different privacy budgets. For the Statlog dataset, there are four sub-figures that represent different ILPG scores with different privacy budgets {0.1, 0.2, 0.3, 0.4}. In each sub-figure, there are two types of ILPG scores. The label VHDP denotes the ILPG score of the vertical heterogeneous DP way with the privacy weight vector $\vec{w} = (\frac{1}{4}, \frac{1}{2}, 1)$, and *DP* denotes the ILPG score of homogeneous DP with the privacy weight vector $\vec{w} = (1, 1, 1)$. The two labels represent different DP ways with different privacy vectors and every colored mark represents the ILPG score of one attribute. In the four sub-figures, we observe that different attributes have different ILPG scores when noise is added. Using our proposed VHDP for every attribute achieves a lower ILPG score than using homogeneous DP. Similar observation is observed for the findings on the Adult dataset.

The results suggest that VHDP is better than homogeneous DP in terms of trade-off level. When we also consider the results from Section 5.2.1, we can conclude that our proposed VHDP method provides strong privacy guarantee with better trade-off level.

6 | RELATED LITERATURE

Since the seminal work of Dwork et al[7] in 2006, the DP model (based on statistical control) has become a popular model in the privacy-preserving literature. Specifically, DP provides a strong security guarantee of individual privacy information (eg, against arbitrary auxiliary knowledge attack) and provides the ability that can be measured. This allows comparisons between different techniques (eg, For a fixed bound on privacy loss, which techniques provide better accuracy? For a fixed accuracy, which techniques provide better privacy?). Existing research relating to DP include mechanism design,[14] applications,[15-18] learning,[19,20] etc.

For the multi-attributes scenario (the setting we work in this paper), there are also a number of existing works utilizing DP, such as those reported in other works,[21-23] evaluating the sensitivity of multi-attributes through reducing dimension or measuring the distance. These approaches generally ensure that every attribute is protected homogeneous together with a global sensitivity without considering partition of the dataset. In other works,[24-31] privacy of multi-attributes data was preserved via vertical partitioning. Such approaches consider how one can reduce the effect of partitioned data from different places on privacy information disclosure between them. However, these approaches do not consider heterogeneous privacy protection in each set of partitioned data. Moreover, these privacy models are not designed to resist an attack utilizing background knowledge and lacks measurement to balance the privacy and data utility.

Alaggan et al[13] recently proposed a DP-based approach to ensure heterogeneous privacy prevention horizontally, and assumed that individuals had different privacy requirements. In other words, different records are protected with different privacy weights according to the requirements, where machine learning techniques are used to assign privacy weights as different privacy demands are selected arbitrarily. The similar idea was applied in the work of Phan et al.[32] However, such an approach does not ensure different privacy protection of attributes.

7 | CONCLUSION

Differential privacy is one influential privacy notion that offers a rigorous and provable privacy guarantee for data learning. However, the conventional DP method can be applied only homogeneously or heterogeneously through horizontal way. With the scenario of multi-attributes data, different attributes potentially influence the privacy differently. This paper has proposed a heterogeneous privacy-preserving mechanism, VHDP, of attributes. Based on information theory, we assign sub-privacy budgets to achieve heterogeneous protection with Laplace mechanism. The experiments indicate that our mechanism is efficient and balance the privacy and utility well.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China under grant 2016QY02D0302 and the National Natural Science Foundation of China (NSFC) under grant 61472148.

ORCID

Xiaofeng Ding https://orcid.org/0000-0001-5054-8515
Hai Jin https://orcid.org/0000-0002-3934-7605

REFERENCES

1. Yang Y, Zheng X, Chang V, Tang C. Semantic keyword searchable proxy re-encryption for postquantum secure cloud storage. *Concurrency Computat Pract Exper*. 2017;29(19).
2. Yang Y, Zheng X, Guo W, Liu X, Chang V. Privacy-preserving fusion of IoT and big data for e-health. *Future Gener Comput Syst*. 2018;86:1437-1455.
3. Yang Y, Zheng X, Guo W, Liu X, Chang V. Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. *Sci China Inf Sci*. 2019;479:567-592.
4. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowledge Based Syst*. 2002;10(05):557-570.
5. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. Paper presented at: 22nd International Conference on Data Engineering; 2006; Atlanta, GA.
6. Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. Paper presented at: 2007 IEEE 23rd International Conference on Data Engineering; 2007; Istanbul, Turkey.
7. Dwork C, McSherry F, Nissim K, Smith AD. Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography*. Berlin, Germany: Springer; 2006:265-284.
8. Shannon CE. A mathematical theory of communication. *Mob Comput Commun Rev*. 2001;5:3-55.
9. Dwork C. A firm foundation for private data analysis. *ACM Communications*. 2011;54:86-95.
10. Fung BCM, Wang K, Yu PS. Top-down specialization for information and privacy preservation. Paper presented at: 21st International Conference on Data Engineering; 2005; Tokyo, Japan.
11. Sadhya D, Chakraborty B, Singh SK. Capturing the effects of attribute based correlation on privacy in micro-databases. In: Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017); 2017; Madrid, Spain.
12. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput Surv*. 2010;42:14:1-14:53.
13. Alaggan M, Gambs S, Kermarrec A-M. Heterogeneous differential privacy. ArXiv:1504.06998 [cs.CR]. 2015.
14. McSherry F, Talwar K. Mechanism design via differential privacy. Paper presented at: 48th Annual IEEE Symposium on Foundations of Computer Science; 2007; Providence, RI.
15. Qardaji WH, Yang W, Li N. Differentially private grids for geospatial data. Paper presented at: 2013 IEEE 29th International Conference on Data Engineering; 2013; Brisbane, Australia.
16. Xu J, Zhang Z, Xiao X, Yang Y, Yu G. Differentially private histogram publication. In: Proceedings of the International Conference on Data Engineering; 2012; Arlington, VA.
17. Zhao J, Jung T, Wang Y, Li X. Achieving differential privacy of data disclosure in the smart grid. Paper presented at: IEEE Conference on Computer Communications; 2014; Toronto, Canada.
18. Bai X, Yao J, Yuan M, Deng K, Xie X, Guan H. Embedding differential privacy in decision tree algorithm with different depths. *Sci China Inf Sci*. 2017;60:082104:1-082104:15.
19. Kairouz P, Oh S, Viswanath P. The composition theorem for differential privacy. *IEEE Trans Inf Theory*. 2017;63:4037-4049.
20. Kusner MJ, Gardner JR, Garnett R, Weinberger KQ. Differentially private Bayesian optimization. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning; 2015; Lille, France.
21. Su S, Tang P, Cheng X, Chen R, Wu Z. Differentially private multi-party high-dimensional data publishing. Paper presented at: 2016 IEEE 32nd International Conference on Data Engineering; 2016; Helsinki, Finland.
22. Chen R, Xiao Q, Zhang Y, Xu J. Differentially private high-dimensional data publication via sampling-based inference. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015; Sydney, Australia.

23. Day W-Y, Li N. Differentially private publishing of high-dimensional data using sensitivity control. In: Proceedings of the ACM Symposium on Information, Computer and Communications Security; 2015; Singapore.
24. Yi X, Zhang Y. Equally contributory privacy-preserving k-means clustering over vertically partitioned data. *Information Systems*. 2013;38:97-107.
25. Lin Z, Jaromczyk JW. Privacy preserving spectral clustering over vertically partitioned data sets. Paper presented at: 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery; 2011; Shanghai, China.
26. Zhu X, Liu M, Xie M. Privacy-preserving affinity propagation clustering over vertically partitioned data. Paper presented at: 2012 4th International Conference on Intelligent Networking and Collaborative Systems; 2012; Bucharest, Romania.
27. Raghuram B, Gyani J. Privacy preserving associative classification on vertically partitioned databases. Paper presented at: 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies; 2012; Ramanathapuram, India.
28. Bansal A, Chen T, Zhong S. Privacy preserving back-propagation neural network learning over arbitrarily partitioned data. *Neural Comput Appl*. 2011;20:143-150.
29. Lin Q, Pei H, Wang K, Zhong P. Privacy-preserving one-class support vector machine with vertically partitioned data. *Multimed Ubiquitous Eng*. 2016;11:199-208.
30. Li L, Lu R, Choo K-KR, Datta A, Shao J. Privacy-preserving-outsourced association rule mining on vertically partitioned databases. *IEEE Trans Inf Forensics Secur*. 2016;11:1847-1861.
31. Vaidya J, Clifton CW. Privacy-preserving kth element score over vertically partitioned data. *IEEE Trans Knowl Data Eng*. 2009;21:253-258.
32. Phan N, Vu MN, Liu Y, Jin R, Dou D, Wu X, Thai MT. Heterogeneous gaussian mechanism: preserving differential privacy in deep learning with provable robustness. arXiv:1906.01444. 2019.

How to cite this article: Xia Y, Zhu T, Ding X, Jin H, Zou D. Heterogeneous differential privacy for vertically partitioned databases. *Concurrency Computat Pract Exper*. 2019;e5607. https://doi.org/10.1002/cpe.5607