DP2-Pub: Differentially Private High-Dimensional Data Publication with Invariant Post Randomization

Honglu Jiang, *Senior Member, IEEE,* Haotian Yu, Xiuzhen Cheng, *Fellow, IEEE,* Jian Pei, *Fellow, IEEE,* Robert Pless, *Member, IEEE,* and Jiguo Yu, *Fellow, IEEE,*

Abstract—A large amount of high-dimensional and heterogeneous data appear in practical applications, which are often published to third parties for data analysis, recommendations, targeted advertising, and reliable predictions. However, publishing these data may disclose personal sensitive information, resulting in an increasing concern on privacy violations. Privacy-preserving data publishing has received considerable attention in recent years. Unfortunately, the differentially private publication of high dimensional data remains a challenging problem. In this paper, we propose a differentially private high-dimensional data publication mechanism (DP2-Pub) that runs in two phases: a Markov-blanket-based attribute clustering phase and an invariant post randomization (PRAM) phase. Specifically, splitting attributes into several low-dimensional clusters with high intra-cluster cohesion and low inter-cluster coupling helps obtain a reasonable allocation of privacy budget, while a double-perturbation mechanism satisfying local differential privacy facilitates an invariant PRAM to ensure no loss of statistical information and thus significantly preserves data utility. We also extend our DP2-Pub mechanism to the scenario with a semi-honest server which satisfies local differential privacy. We conduct extensive experiments on four real-world datasets and the experimental results demonstrate that our mechanism can significantly improve the data utility of the published data while satisfying differential privacy.

Index Terms—High-dimensional data, differential privacy, Bayesian network, Markov-blanket, invariant PRAM.

1 Introduction

The rapid development of information technology has opened up the era of big data. Collecting and publishing an unprecedented amount of data, as well as mining data correlations and generating insights have become an important component of social statistical research [1]. A large amount of high-dimensional and heterogeneous data appear in various applications, which are often published to third parties for data analysis, recommendations, targeted advertisements, and reliable predictions. Examples include healthcare data, social networking data, Internet of Things data (i.e., IoT device monitoring data, location data, trajectory data), financial market data (i.e., electronics commercial data, credit card data), which can be used to dig out valuable information hidden behind the massive data for modern life. However, publishing these data may disclose personal sensitive information, resulting in an increasing concern of privacy violations. Privacy-preserving data publishing (PPDP) has gained significant attentions in recent years as a promising approach for information sharing while preserving data privacy [2].

Generally speaking, commonly used approaches for PPDP can be characterized into three categories: encryption technology [3], -anonymity [4] and its derivative approaches (-diversity [5], -closeness [6]), and differential privacy [7]. Differential privacy has gradually become the *de facto* standard privacy definition and provides a strong privacy guarantee. It rests on a sound mathematical foundation with a formal definition and rigorous proof while making the assumption that an attacker has the maximum background knowledge.

However, the differentially private publication of high dimensional data remains a challenging problem – it suffers from the "Curse of High-Dimensionality" [8], that is, when the dimensionality increases, the complexity and cost of multi-dimensional data processing and analysis increases exponentially. Specifically, this curse is manifested in two aspects: first, since the high-dimensional data space is usually sparse, high dimensions and large attribute domains lead to a low "Signal-to-Noise Ratio" and low data utility; second, complex correlations exist between high-dimensional data, therefore the change of a single record may have a great impact on query results, leading to increased sensitivity.

To address these challenges, an effective way is to decompose high-dimensional data into a set of low-dimensional marginal tables along with inferring the joint distributions of the data, thus generating a synthetic dataset.

H. Jiang is with the Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, USA, and the Department of Computer Science, The George Washington University, Washington, DC 20052, USA. E-mail: jiangh34@miamioh.edu.
H. Yu is with the Department of Data Analytics, The George Washington University, Washington, DC 20052, USA. E-mail: yuxx6789@gwu.edu.
X. Cheng is with the Department of Computer Science, The George Washington University, Washington, DC 20052, USA, and the School of Computer Science and Technology, Shandong University, Qingdao 266510, China. E-mail: xzcheng@sdu.edu.cn.
J. Pei is with the Departments of Computer Science, Biostatistics and Bioinformatics, and Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. E-mail: j.pei@duke.edu.
R. Pless is with the Department of Computer Science, The George Washington University, Washington, DC 20052, USA. E-mail: pless@gwu.edu.
J. Yu (Corresponding Author) is with Big Data Institute, Qilu University of Technology, Jinan, 250353, P.R. China. E-mail: jiguoyu@sina.com.

Manuscript received; revised.

A representative solution is PrivBayes [9], which constructs a Bayesian network to model the data correlations and conditional probability distributions, allowing one to approximate the distributions of the original data using a set of low-dimensional marginal distributions. However, such an approach suffers from poor data utility and high communication cost, since too much noise is added when there are too many attribute pairs resulting in unreliable conditional probabilities. Moreover, most approaches generally ignore the different roles a dimension may play for a specific query – one dimension may be more important than another for a particular query. Additionally, one dimension may release more information than another if the same amount of noise is added; thus evenly allocating the total privacy budget to each dimension degrades the performance.

In this paper, we provide a two-phase mechanism (DP2-Pub) consisting of a Markov-blanket-based learning process and an invariant post randomization (PRAM) process satisfying local differential privacy to overcome the above difficulties. Our contributions can be summarized as follows:

- To capture the dependencies between the attributes in the dataset, we resort to differentially private Bayesian network construction, employing the exponential mechanism to attribute pairs using the mutual information as the score function.
- We propose the procedure of attribute clustering with a Markov blanket learning algorithm based on the constructed Bayesian network. Our most fundamental purpose is to split attributes into several low-dimensional clusters with high intra-cluster cohesion and low inter-cluster coupling, thus obtaining a reasonable allocation of privacy budget determined by the conditional independence among attributes and the importance of each cluster.
- Invariant PRAM is an important perturbation technique for privacy protection, which transforms each record stochastically in a dataset using delicately pre-selected probabilities. It ensures no loss of statistical information, thus can significantly preserve data utility. Motivated by this, we provide a double-perturbation mechanism to achieve invariant PRAM and differential privacy for two-valued and multivalued attributes, then apply it to each attribute cluster. Resorting to the randomized mapping based post-processing property for differential privacy, we prove that the proposed double-perturbation mechanism satisfies differential privacy.
- To tackle the data privacy preservation problem for the scenario where each individual contributes a single data record to a semi-honest server, we extend our DP2-Pub mechanism to handle the high-dimensional data publication in a local-differential-privacy manner, in which each user locally perturbs its data satisfying local differential privacy, then the server conducts all the operations including attribute clustering and post randomization over the privatized data.
- We evaluate the performance of data utility on four real-world datasets from two aspects, the total variation distance between the original dataset and the perturbed dataset and the classification error rate of SVM classification on the perturbed dataset. Experimental results indicate that our approach can obtain higher data utility of the published data compared with the state-of-the-art.

The rest of this paper is organized as follows. We provide a literature review in Section 2. Section 3 formulates our problem and presents necessary background knowledge on Bayesian network, differential privacy, random response and post randomization. In Section 4, we propose our DP2-Pub mechanism by detailing the constructions of differentially private Bayesian network, attribute clustering, and invariant PRAM. Comprehensive experimental studies on four real-world datasets are presented in Section 6. Section 7 concludes the paper with a future research discussion.

2 Related Work

Various differentially private mechanisms for high-dimensional data publications have been proposed in recent years. In this section, we briefly review the most relevant works from two perspectives: under centralized setting or distributed setting, and discuss how our work differs from the existing ones.

2.1 Private Mechanisms Under Centralized setting

A powerful approach of dimensionality reduction is the Bayesian network model proposed in [9], in which Zhang *et al.* developed a differentially private scheme PrivBayes for publishing high-dimensional data. PrivBayes first constructs a Bayesian network to approximate the distribution of the original dataset, then adds noise into each marginal of the Bayesian network to guarantee differential privacy, next constructs an approximate distribution of the original dataset, and finally samples the tuples from the approximate distribution to construct a synthetic dataset.

Researchers also have developed sampling techniques to support differentially private high-dimensional data publications. In [10], Chen *et al.* provided a solution to protect the joint distribution of the dimensions in a high-dimensional dataset compared with PrivBayes. They first established a robust sampling-based approach to investigate the dependencies over all attributes for constructing a dependence graph, then applied a junction tree algorithm to provide an inference mechanism for deriving the joint data distribution. Both [9] and [10] constructed a dependency graph and generated a differentially private marginal table to enforce consistency constraints over all marginals, during which they evenly split the privacy budget into portions, each being used for a pair of attributes.

In [11], Li *et al.* proposed a differentially private data synthetization technique called DPCopula using Copula functions to handle multi-dimensional data. In [12], Xu *et al.* developed a high-dimensional data publishing algorithm under differential privacy to optimize the utility by first projecting a -dimensional vector of user's attributes into a lower -dimensional space using a random projection, then adding Gaussian noise to each resultant vector to obtain a synthetic dataset. The authors represent each user's feature attributes as a -dimensional vector and ignore the different

roles a dimension may play for a specific query and the correlation between different attributes.

2.2 Private Mechanisms Under Distributed setting

The approaches mentioned above mainly consider centralized scenarios. Some efforts have also been devoted to differentially private high-dimensional data publications under distributed setting. Based on PrivBayes, Cheng *et al.* [13] considered a multi-party setting from multiple data owners and proposed a differentially private sequential update of the Bayesian network (DP-SUBN) approach, allowing the parties to collaboratively identify the Bayesian network that best approximates the joint distribution of the integrated dataset. Wang *et al.* [14] introduced a framework with a simple and generic aggregation and decoding technique. This framework can analyze, generalize and optimize several local differential privacy protocols [15–17] for frequency estimation. In [8], Ren *et al.* proposed a solution LoPub to realize high-dimensional data publication with local differential privacy in crowdsourced data publication systems. LoPub can first learn from the distributed data records to build correlations and joint distributions of attributes, then synthesize an approximate dataset achieving a good compromise between local differential privacy and data utility. In [18], Ju *et al.* also considered the high-dimensional data publication problem under local differential privacy in the crowdsourced-sensing system. They proposed an aggregation and publication mechanism which provides local privacy guarantees for crowd-sensing users, approximates the statistical characteristics of high-dimensional perception data and publishes synthetic data. Wang *et al.* [19] proposed two mechanisms for collecting and analyzing users' private data under local differential privacy, which can collect multidimensional data with both numerical and categorical attributes. In [20], Domingo-Ferrer developed several random-response-based complementary approaches for multi-dimensional data preservation. In [21], Takagi *et al.* presented a privacy-preserving phased generative model (P3GM) for high-dimensional data, which employs a two-phase learning process for training the model to increase the robustness to the differential privacy constraint.

In light of the above analysis, the following aspects distinguish our work from the existing approaches. First, since the sensitivity of distinct dimensions are different and evenly allocating the total privacy budget to each dimension cannot obtain good performance, we consider the privacy budget allocation problem to realize attribute clustering with a reasonable allocation of privacy budget. Second, we design a double-perturbation mechanism to achieve invariant PRAM instead of generating noisy conditional distributions of the Bayesian network, then apply it to each attribute cluster, which can significantly improve the data utility while satisfying local differential privacy.

3 Problem Formulation and Preliminaries

3.1 Problem Formulation

In this paper, we consider the following problem: a data server collects data containing a vast amount of individual information and aims to release an approximate dataset to third parties for their uses such as data analysis and recommendations. Let be the dataset, be the total number of records, and $\mathcal{A}$ be the set of unique attributes. Assume that all attribute values are categorical as one can always discretize all numerical data. The domain of an attribute is denoted by , whose size is .

3.2 Bayesian Network

A Bayesian network is a type of probabilistic graphical model that approximately describes the joint distribution over a set of variables by specifying their conditional independence [22]. More specifically, a Bayesian network is a directed acyclic graph (DAG) whose nodes represent attribute variables and edges model the direct dependence among attributes. Formally speaking, a Bayesian network $\mathcal{N}$ over $\mathcal{A}$ (the set of attributes in) is defined as a set of attribute-parent (AP) pairs, \ \ , where each AP contains a unique attribute and all its parent nodes in $\mathcal{N}$. If the maximum size of any parent set in $\mathcal{N}$ is , we define $\mathcal{N}$ to be a -degree Bayesian network. Let $\mathcal{A}$ denote the joint distribution over all attributes in . A Bayesian network $\mathcal{N}$ defines a way to approximate $\mathcal{A}$ with conditional distributions , that is, $_{\mathcal{N}}\ \mathcal{A}$. If $\mathcal{N}$ accurately captures the dependencies between the attributes in $\mathcal{A}$, $_{\mathcal{N}}\ \mathcal{A}$ can be a good approximation to $\mathcal{A}$. Moreover, the computation of $_{\mathcal{N}}\ \mathcal{A}$ can be efficient and simple if the degree of $\mathcal{N}$ is small. Figure 1 illustrates the Bayesian network over a set of five attributes, namely, *age, gender, exposure to toxins, smoking*, and *cancer*. Table 1 shows the AP pairs in the sample Bayesian network.

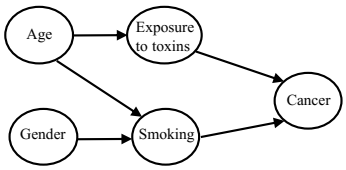


Fig. 1. A Bayesian network over five attributes.

TABLE 1
The attribute-parent pairs in the Bayesian network shown in Fig. 1

Attribute A_i	Π_i
Age	
Gender	
Exposure to toxins	{Age}
Smoking	{Age, Gender}
Cancer	{Exposure to toxins, Smoking}

Given a dataset , our goal is to construct a -degree Bayesian network $\mathcal{N}$ which provides an accurate approximation to the full distribution of . That is, $_{\mathcal{N}}\ \mathcal{A}$ should be close to $\mathcal{A}$. As -divergence [23] is commonly used as a measure of the similarity between a probability distribution and a candidate (estimated) distribution, in this paper, we adopt the -divergence of $_{\mathcal{N}}\ \mathcal{A}$ and $\mathcal{A}$ to measure the difference between these two probability distributions:

\ $\mathcal{A}$ $_{\mathcal{N}}\ \mathcal{A}$ \ \ \$\mathcal{A}$

where \ denotes the entropy of the random variable ,

\

\

and \ denotes the mutual information between the two variables:

\

\

Here, is the joint distribution of and , and are the marginal distributions of and , respectively, and \$\mathcal{A}$ is the joint entropy of all attribute variables in $\mathcal{A}$, which is defined as:

\$\mathcal{A}$ \

\ \

Therefore, learning a Bayesian network is to find $\mathcal{N}$ from with the minimum \ $\mathcal{A}$ $\mathcal{N}$ $\mathcal{A}$. The construction of $\mathcal{N}$ can be modeled as choosing a parent set for each attribute to maximize \ since \ \$\mathcal{A}$ is fixed once the dataset is given.

3.3 Differential Privacy

Differential privacy (DP) has become the *de facto* standard of privacy preservation, which ensures that query results of a dataset are insensitive to the change of a single record. Differential privacy is defined based on the neighboring datasets and , where differs from by only one record:

Definition 1 (Differential privacy [24])**.** *A randomized algorithm is -differentially private if for any pair of neighboring datasets and , and for all sets of possible outputs, we have*

\ \

where is often a small positive real number.

The smaller the , the higher the level of privacy preservation. A smaller provides greater privacy preservation at the cost of lower data accuracy with more added noise. Differential privacy can be achieved by two best known mechanisms, namely the *Laplace mechanism* [24] and the *exponential mechanism* [25]. We provide the formal definition of the exponential mechanism as follows:

Definition 2 (Exponential Mechanism [25])**.** *Given a random algorithm with the input dataset and the output entity object , where is the output range. Let \ be the utility function and be the global sensitivity of function \ . If algorithm selects and outputs from at a probability proportional to \——— , then is -differentially private.*

DP techniques implicitly assume a trusted third party to collect data and thus can hardly be applied to the case where a server is not reliable. Therefore *local differential privacy* (LDP) emerges, in which each user independently and locally conducts data perturbation. The formal definition of LDP can be shown as follows:

Definition 3 (Local Differential Privacy [26])**.** *Consider records. A privacy algorithm with domain \ and range \ satisfies the -local differential privacy if obtains the same output result \ \ on any two records and \ \ with*

\ \

Local differential privacy ensures the similarity between the output results of any two records. Random response (RR) [27] is currently the most widely used technique for achieving local differential privacy.

3.4 Random Response and Post Randomization

Random response (RR) is a technique developed in social science to collect statistical data about individuals' sensitive information. Its main idea is to provide data privacy protection by making use of the uncertainty of responses to sensitive questions. Privacy comes from the randomness of the answers while accuracy comes from the noise generation procedure [28].

Post randomization (PRAM) is another important perturbation technique for privacy protection, which stochastically transforms each record in a dataset using pre-selected probabilities. For a random variable with categories , let , and \ . The basic idea of PRAM is to select a transition probability matrix \ with for . Then the original category is changed to with probability . Let denote the transformed variable. We have \ , , and \ .

Mathematically, PRAM is equivalent to RR. Therefore many mathematical results developed for RR such as the local differential privacy guarantee can be applied to PRAM [29]. In this paper, we employ PRAM for the case when a trusted data server is available (Section 4), where all the data can be processed at the server to maintain differential privacy, and RR for the case when the server is semi-honest, in which case local differential privacy is adopted for collecting data from each user to the server (Section 5).

4 DP2-Pub With a Trusted Server

In this section, we propose a novel **d**ifferentially **p**rivate high-dimensional data **pub**lication mechanism based on a **d**ouble-**p**erturbation process, namely DP2-Pub, assuming the availability of a trusted server that can access the original data. We first present an overview on DP2-Pub, then detail its modules in the following subsections.

4.1 Overview

Figure 2 illustrates the main procedure of DP2-Pub, which runs in two phases of attribute clustering and data randomization, with both being performed by the trusted server. Since both phases require access to the original dataset, we divide the total privacy budget into two portions with being used for the first phase and for the second phase, and demonstrate that the two phases are both differentially private.

1. Bayesian Network and Attribute Clustering. To learn the correlations between different attribute variables, we adopt the approach of constructing a differentially private

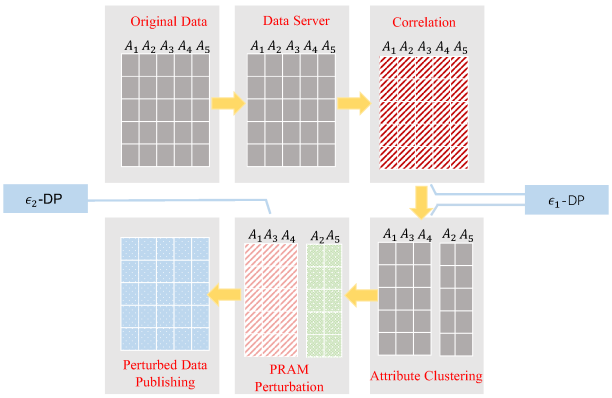


Fig. 2. Overview of DP2-Pub.

Algorithm 1: DP-Bayesian Network Construction

Input: Dataset D; k, the degree of the Bayesian network; and the set of attributes $\mathcal{A}$
Output: Bayesian network $\mathcal{N}$

1 Initialize $\mathcal{N}$ and V ;
2 Randomly select an attribute as A , add A to V, add A , to $\mathcal{N}$;
3 **for** i 2 *to* d **do**
4 initialize Ψ ;
5 for each A $\mathcal{A}$ V and each Π , add A , Π to Ψ;
// denotes the set of all subsets of V with size of min k, V
6 Select an AP pair A , Π from Ψ at a probability proportional to exp ; //Δ denotes the sensitivity of the mutual information function.
7 Add A to V and A , Π to $\mathcal{N}$;
8 **return** $\mathcal{N}$

Bayesian network through the exponential mechanism presented in [9]. Based on the constructed Bayesian network, we propose the procedure of attribute clustering using the Markov blanket model to achieve high intra-cluster cohesion and low inter-cluster coupling. Each cluster is composed of a cluster head and its Markov blanket members, thus the attribute set can be divided into a number of disjoint clusters denoted as . Our most fundamental purpose is to realize attribute clustering, and then to obtain a reasonable allocation of privacy budget for each cluster based on its importance.

2. Data Randomization. We propose a detailed double-perturbation mechanism to achieve invariant post randomization and differential privacy, then apply it to each attribute cluster. Note that PRAM is an important technique for data perturbation, and that a PRAM is invariant if the transition probability matrix satisfies (except for the identity matrix $\mathbb{I}$). The appealing advantage of an invariant PRAM lies in that there is no loss of statistical information and thus can significantly preserve data utility. The key point to construct an invariant PRAM is to delicately solve satisfying . We design a double-perturbation mechanism to achieve the invariant property of PRAM for two-valued and multivalued attributes.

4.2 Differentially Private Bayesian Network Construction

In this section, we adopt the algorithm proposed in [9] to construct our Bayesian network in a differentially private manner, which employs the exponential mechanism to select , using the mutual information as the score function. For completeness, we present the algorithm in Algorithm 1.

At the beginning of Algorithm 1, we initialize the Bayesian network $\mathcal{N}$ to be an empty set. Let denote the set containing the attributes whose parent sets have been fixed and the initial set of is empty (Line 1). Then we randomly select an attribute from the attributes as the initial node and set its parent set empty (Line). For each , its AP pair is selected in a differentially private manner by the exponential mechanism (Lines -), of which the mutual information is taken as the score function, and is its global sensitivity. Algorithm 1 ensures that each invocation of the exponential mechanism satisfies -differential privacy, and the exponential mechanism is invoked times, so the construction of $\mathcal{N}$ is -differentially private based on DP's composability property (Lines). We adopt the calculation of the sensitivity of the mutual information in [9], which is shown as follows:

– — — if or is binary,
– — — — otherwise. (1)

4.3 Attribute Clustering

Given a Bayesian network, the Markov blanket of an attribute variable can be intuitively represented as the set of parent nodes and child nodes of as well as the set of parent nodes of 's child nodes, which can be formalized as follows:

We propose a procedure of attribute clustering shown in Algorithm 2. First, we initialize the set to include all the attributes of $\mathcal{N}$. Then we randomly select an attribute , add and into a cluster, and delete all these attributes from . Repeat this procedure until is empty. Each cluster is composed of a cluster head (an attribute variable) and its Markov blanket members. Thus, the attribute set can be divided into a number of disjoint clusters, which can be denoted as .

Algorithm 2: Attribute Clustering

Input: Bayesian Network $\mathcal{N}$
Output: Cluster CL , CL , CL

1 S=Set of all attributes of $\mathcal{N}$;
2 i=0;
3 **while** S **do**
4 i=i+1;
5 Randomly select the attribute x in S and let CL MB x x ;
6 S S CL ;
7 **return** CL , CL , CL

The key to overcome the curse of dimensionality is to decompose high-dimensional data into a set of low-dimensional data based on the conditional independences of the data. Bayesian network and Markov blanket are the

most widely used graphical models for identifying a minimal set of attributes with strong correlations. Specifically, for any attribute variable A_i in the Bayesian network, its Markov blanket is the set of attributes which are strongly corelated to A_i, while the attributes not in A_i's Markov blanket are loosely correlated with A_i or even conditionally independent of A_i. Therefore, our clustering algorithm yields attribute clusters with high intra-cluster correlation (cohesion) and low inter-cluster coupling which can improve the accuracy of the estimated joint distribution of the data. Note that the input of Algorithm 2 is the differentially private Bayesian network constructed from Algorithm 1, which guarantees that the operation of attribute clustering does not break differential privacy.

According to the attribute clustering process, a reasonable allocation of the privacy budget for the next data randomization phase is determined by the conditional independence among the attributes in a cluster and the importance of the cluster based on the probability distributions over the dataset. Thus we define the importance factor (CIF) of each cluster $CL_i, 1 \le i \le t$, in (2), which measures the importance of each cluster. The higher the CIF, the more important the cluster.

$$CIF(CL_i) = \frac{\sum_{A_j \in CL_i} H(A_j)}{\sum_{k=1}^{d} H(A_k)} \tag{2}$$

Based on the CIF, one can allocate a privacy budget to each cluster, following the principle stating that the smaller the privacy budget, the higher the level of privacy preservation. Therefore, we define the privacy budget coefficient (PBC) for each cluster CL_i as follows:

$$PBC(CL_i) = \frac{\frac{1}{CIF(CL_i)}}{\sum_{j=1}^{t} \frac{1}{CIF(CL_j)}} \tag{3}$$

Note that (3) conducts a normalization of PBC so that it falls into the $[0, 1]$ interval. As the privacy budget of the data randomization is ϵ_2, which will be carried out on each cluster at the server, the privacy budget allocated for cluster CL_i is $PBC(CL_i) \cdot \epsilon_2$.

4.4 Invariant PRAM

The characteristic of an invariant PRAM lies in that the transition probability matrix P satisfying $P\vec{\pi} = \vec{\pi}$. In this section, we propose a detailed double-perturbation scheme to achieve invariant PRAM and differential privacy, which is suitable for categorical attributes. The main idea of our approach is to compute P via double-perturbation, as shown in Figure 3. For an attribute variable X, let X_1 denote the perturbed variable after the first perturbation, and X_2 denote the one after the second perturbation. We first construct a transition probability matrix $Q = (q_{ij})$ satisfying differential privacy and conduct the first perturbation on the attribute variable X according Q. Then we compute the estimate of π based on the perturbed data X_1, denoted as $\hat{\vec{\pi}}$, construct the transition probability matrix $\tilde{Q} = (\tilde{q}_{ij})$ for the second perturbation according to a specific rule to achieve $\tilde{Q} \cdot Q \cdot \vec{\pi} = \vec{\pi}$, and finally carry out the second perturbation on X_1 based on $\tilde{Q}$ to obtain X_2. More specifically, the rule of constructing $\tilde{Q}$ is to set $\tilde{q}_{ji} = Pr(X = c_i | X_1 = c_j)$, where $\tilde{q}_{ji}$ denotes the probability of X_1 being changed from category c_j to c_i in the second perturbation. In other words, $\tilde{Q}$ can be considered as the inverse of Q while ensuring that $\tilde{Q}$ is also a transition probability matrix. Thus, the double-perturbation is an invariant PRAM with $P = \tilde{Q} \cdot Q$, and it satisfies differential privacy since the first perturbation satisfies differential privacy and the second perturbation is a randomized mapping of the first one.

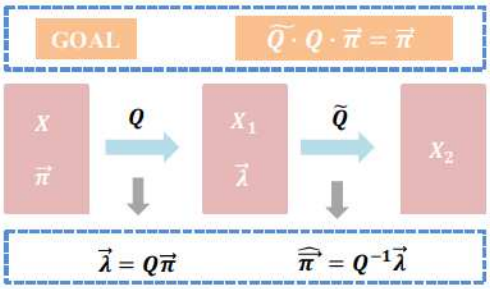


Fig. 3. Process of Double Perturbation.

The attribute variables can be either two-valued or multivalued. For better elaboration, we detail the construction of the transition probability matrix for a two-valued attribute first and then extend the procedure to multivalued attributes. Following that, we apply the construction to compound variables and present the analysis on the privacy guarantee.

4.4.1 Two-Valued Attributes

We start from the case of a categorical attribute variable X with only two possible values of c_1 and c_2. Let $\pi_1 = Pr[X = c_1]$ and $\pi_2 = Pr[X = c_2]$; denoted by $\vec{\pi} = (\pi_1, \pi_2)^{\mathrm{T}}$. The data randomization process is conducted in a way that c_1 or c_2 either remains unchanged with probability q, or is changed to the other value with the probability of $1 - q$; that is $q_{11} = q_{22} = q$ and $q_{12} = q_{21} = 1 - q$. Thus, the transition probability matrix Q of the two-valued variable X is a 2×2 matrix, which can be shown as follows:

$$Q = \begin{bmatrix} q & 1-q \\ 1-q & q \end{bmatrix}$$

Let $\lambda_1 = Pr[X_1 = c_1]$ and $\lambda_2 = Pr[X_1 = c_2]$; set $\vec{\lambda} = (\lambda_1, \lambda_2)^{\mathrm{T}}$. Note that the transition probability matrix satisfies $\vec{\lambda} = Q\vec{\pi}$.

In this setting, let $q = \frac{e^{\epsilon}}{1+e^{\epsilon}}$. Then the local differential privacy can be satisfied as:

$$\frac{Pr(X_1 = c_1 | X = c_1)}{Pr(X_1 = c_1 | X = c_2)} \le \frac{q}{1-q} \le e^{\epsilon}$$

As mentioned earlier, our mechanism achieves invariant post randomization to preserve the statistical information and data utility to the greatest possible extent, which first adopts transition probability matrix Q on the original attribute variable X, estimates the probability distribution of the variable X based on the perturbed data after the first perturbation, and constructs the transition probability matrix $\tilde{Q}$ for the second perturbation according to the transition probability matrix Q and the probability distribution

of the perturbed data . The advantage of this double-perturbation mechanism lies in that there is no need to know the probability distribution of the original data in advance– we actually do not know the probability distribution of –the transition probability matrix of the original data is thus constructed adaptively. After obtaining the perturbed data with , we can obtain the estimate of the original attribute variable distribution:

(4)

Then we compute the transition probability of each variable for the second perturbation as follows:

Accordingly, we obtain the transition probability matrix for the second perturbation:

Therefore, to obtain the invariant PRAMed data of the attribute variable , we apply to the perturbed data during the second perturbation. These two phases of data perturbation with and successfully realize an invariant PRAM with , where can be considered as the inverse of while ensuring that is also a transition probability matrix.

4.4.2 Multivalued Attributes

The perturbation of the multivalued attributes is similar to that of the two-valued one. We consider a categorical random variable with possible values . Let and \ . Given that belongs to category , it either remains unchanged with probability , or is changed uniformly at random with the probability of —— to one of the other categories. That is, the transition probability matrix is a one, which can be formalized as follows:

To satisfy local differential privacy, we set ———; then can be denoted as:

The PRAMed variable can be denoted as after applying to . Correspondingly, let , \ . In this setting, the local differential privacy condition can be satisfied as:

We can compute the estimated of the original attribute variable as

Then the elements of the transition probability matrix for the second perturbation can be computed as:

It can be observed that , which satisfies the property of a transition probability matrix. We take as an example:

After applying to the perturbed data in the second perturbation, we obtain the invariant PRAM result of the attribute variable .

4.4.3 Compound Variables

Since an attribute cluster may include more than one two-valued or multivalued attribute variables which are strongly correlated, one can treat all these variables as a compound one. Thus an invariant PRAM for compound variables [29] is needed, which first computes the transition probability matrix for each attribute variable, then computes the transition probability matrix for the compound one. For example, for two categorical variables with categories and with categories, we may first compute the invariant PRAM transition probability matrix of and , denoted as \ and \ , respectively. Then the combination of and can

[

\ ]

Similarly, when there exist three categorical variables with respectively categories, we may first

compute the invariant PRAM transition probability matrix of , and , denoted as \ , \ and \ . Then the combination of , and can be regarded as a compound variable with categories, whose transition probability is the Kronecker product of , and , which is a \ \ matrix.

As mentioned earlier, when applying the proposed invariant PRAM to each cluster, we need to allocate privacy budget \ to cluster . If a cluster includes more than one attribute variable, we first compute the transition probability matrix for each attribute variable with uniformly allocated privacy budget ——\——, then compute the transition probability matrix for the compound variable.

4.5 Privacy Analysis

As discussed in Section 4.2, Algorithm 1 satisfies -differential privacy, i.e., the procedure of Bayesian network construction satisfies differential privacy. The procedure of attribute clustering just simply cluster the attributes based on the constructed Bayesian network, which does not disclose more information. Therefore, one can say that the first phase of DP2-Pub, i.e., Bayesian network construction and attribute clustering, satisfies -differential privacy.

Now we analyze the second phase, i.e., the phase of data randomization. According to [28], differential privacy is resistant to any randomized mapping of differentially private results. More specifically, with randomized mapping, a data analyst cannot make the output of a differentially private algorithm less differentially private without any additional knowledge about the private dataset [28]. That is, if an algorithm is differentially private, simply conducting randomized mapping on the output of the algorithm without any additional knowledge does not leak any extra private information, which has been proved by the following Post-Processing Proposition [28]:

Proposition 1. *(Post-Processing [28]) Let → be a randomized algorithm that is -differentially private. Let → be an arbitrary randomized mapping. Then → is -differentially private.*

Theorem 1. *The double-perturbation of Invariant PRAM satisfies -differential privacy.*

Proof. The first perturbation of each attribute variable is a post randomization satisfying local differential privacy for both two-valued and multivalued variables:

\ ——

\

\ ——

\ \

of which and are determined by the privacy budget allocated for each attribute variable in a cluster. According to Proposition 1, the second random perturbation of our invariant PRAM mechanism can be considered as a randomized mapping of the differentially private algorithm output, that is, a randomized mapping based post-processing of differential privacy.

Since each cluster is \ -differentially private, the clusters can be regarded as a -dimensional dataset achieving -differential privacy according to the sequential composition theorem [30]. □

Accordingly, one can obtain the following theorem.

Theorem 2. *The DP2-Pub satisfies \ -differential privacy according to sequential composition theorem [30].*

5 DP2-Pub With a Semi-honest Server

The emergence of Internet of Things (IoT) has changed people's daily life and the way the world learns, where mobile devices, home appliances, transportation facilities and crowd sensors can all be used as data acquisition equipment in IoT. It provides a platform for the seamless communication between smart devices and sensors in a smart environment and allows information sharing across platforms. IoT devices and the generated data can reveal personal information of the users including their behaviors and preferences [31]. Despite the benefits of the IoT, it raises privacy concerns of the sheer amount of data. Most of existing privacy-preserving data publishing mechanisms focus on the processing of the collected data with a trustful central server. However, what is stored in the server is unprotected while the central server is vulnerable to internal attacks or single-point attacks; even the server itself may not be trustworthy – it is generally semi-honest, i.e., honest-but-curious, which faithfully follows the protocol but tries its best to infer as much knowledge as possible. Moreover, the data or updates (under federated learning framework) held by the resource-constrained devices can be easily observed or analyzed, which may pose a threat to the privacy protection of participating devices and ultimately discourages participation in the distributed model.

Therefore, in this section, we extend our DP2-Pub mechanism to consider a semi-honest server. A number of users generate multi-dimensional data records, then send them to a server who intends to release an approximate dataset to third-parties for various applications. Formally, each user contributes a data record constituting a dataset , where denotes the data record of user and is the total number of records/users.

Figure 4 illustrates the main procedure of DP2-Pub with a semi-honest server, which includes three main steps: privacy preservation of local data satisfying local differential privacy, Markov-blanket-based cluster learning based on Bayesian network, and the PRAM perturbation on the private data. Both the attribute clustering and PRAM perturbation are conducted at the data server, while the local differential privacy protection is performed by each user. Although the data server is semi-honest, it can only access the private data processed by each user.

We first propose a local randomization using RR on each user's data making it satisfy LDP, then the sanitized data is sent to and aggregated at the central server. Each user has a -dimensional data record , and the perturbation process is conducted on each dimension with the privacy budget .

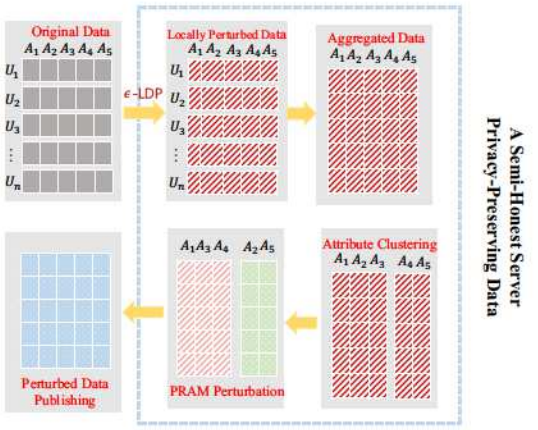


Fig. 4. Overview of DP2-Pub with a Semi-honest Server.

If u_j^i is the value of a two-valued attribute, it is randomly flipped according to the following rule in RR:

$$u_j^i = \begin{cases} u_j^i, & \text{with probability of } q = \frac{e^{\epsilon'}}{1+e^{\epsilon'}} \\ 1 - u_j^i, & \text{with probability of } \frac{1}{1+e^{\epsilon'}} \end{cases}$$

If u_j^i is the value of a multivalued attribute with s possible values $c_1, c_2, \cdots c_s$, it is randomly flipped according to the following rule in RR:

$$u_j^i = \begin{cases} u_j^i, & \text{with probability of } q_{ss} = \frac{e^{\epsilon_i}}{s-1+e^{\epsilon'}} \\ c_k \neq u_j^i, & \text{with probability of } \frac{1}{s-1+e^{\epsilon'}} \end{cases}$$

of which $k = 1, 2, \cdots, s$.

After receiving the noisy data from each user, the server computes the marginal probability distribution $\tilde{\lambda}$, estimates the original distribution $\hat{\tilde{\pi}}$, and then calculates Q for each attribute variable according to the methods presented in Sections 4.4.1 and 4.4.2. Then it constructs a Bayesian network and conducts attribute clustering on the aggregated data to learn the correlations between different attribute variables. The processes of Bayesian network construction and attribute clustering are similar to those in Sections 4.2 and 4.3 except for a few minor changes: replace Line 6 of Algorithm 1 with a procedure that selects (A_i, Π_i) with the largest $I(A, \Pi)$, since the process of the Bayesian network construction does not need to satisfy differential privacy, as the aggregated data at the server is already differentially private (guaranteed by local differential privacy). To further improve accuracy, Line 5 of Algorithm 2 can be replaced by "Select the attribute x with the maximal entropy in S".

Next the server calculates $\tilde{Q}$ for each attribute variable based on $\hat{\tilde{\pi}}$ and Q. Then the server conducts the randomized perturbation by applying $\tilde{Q}$ on the aggregated data to achieve invariant PRAM. According to the attribute clustering, each cluster may include more than one attribute variable which are strongly correlated. Thus we compute the transition probability matrix of the compound variable following the procedure presented in Section 4.4.3.

Theorem 3. *The DP2-Pub with a semi-honest server satisfies ϵ-local differential privacy.*

Proof. Each user perturbs its data record individually with the help of random response to get the privatized data, which provides local differential privacy. The operations of the server are all conducted on the privacy-preserved data, and the PRAM perturbation can be considered as a randomized mapping (post processing) without breaking differential privacy. Therefore, the DP2-Pub mechanism with a semi-honest server is differentially private with privacy budget ϵ. □

6 Experimental Evaluations

In this section, we conduct extensive experiments to demonstrate the performance of our DP2-Pub mechanism and compare it with two benchmark approaches, PrivBayes [9] and DPPro [12], on four real-world datasets of NLTCS [32], ACS [33], BR2000 [33] and Adult [34]. The data utility is evaluated and analyzed from two aspects, namely the total variation distance between the original dataset and the perturbed dataset, and the classification error rate of the SVM classification on the perturbed datasets.

6.1 Experimental Settings

6.1.1 Datasets

We make use of four real-world datasets in our experiments: NLTCS [32] consists of records of 21574 individuals participated in the National Long Term Care Survey, and each record has 16 attributes; ACS [33] includes 47461 records of personal information from the 2013 and 2014 ACS sample sets in IPUMS-USA, where each record has 23 attributes; BR2000 [33] consists of 38000 census records with 14 attributes collected from Brazil in the year 2000; and Adult [34] contains personal information such as gender, salary, and education level of 45222 records extracted from the 1994 US Census, where each record has 15 attributes. The first two datasets only contain binary attribute values while the last two possess continuous as well as categorical attributes with multiple values. We summarize the statistics of these datasets in Table 2.

TABLE 2
Data Statistics

Dataset	Cardinality	Dimensionality	Domain size
NLTCS	21574	16	$\approx 2^{16}$
ACS	47461	23	$\approx 2^{23}$
BR2000	38000	14	$\approx 2^{32}$
Adult	45222	15	$\approx 2^{52}$

6.1.2 Evaluation Metrics

We consider two tasks to evaluate the performance of DP2-Pub. The first task is to study the accuracy of α-way marginals of the perturbed datasets. We evaluate the α-way marginals of the four datasets by adopting the *total variation distance* [23] between the noisy marginal distribution and that of the original datasets, which is shown in Eq. (5).

$$AVD(X, Z) = \frac{1}{2}\|P_X - P_Z\|_1 = \frac{1}{2}\sum_{w\in\Omega}|P_X(w) - P_Z(w)| \quad (5)$$

where Ω is the domain of the probability variable X and Z; P_X and P_Z are the probability distributions of the original attribute variable X and the perturbed one Z, respectively.

Then we compute the average results of the total variation distance over all -way marginals as the final result – a lower distance implies a better utility. More specifically, in our experiments, we evaluate the -way and -way marginals on binary datasets NLTCS and ACS, and -way and -way marginals on BR2000 and Adult, since the domain size of BR2000 and Adult are prohibitively large leading to very complex joint distributions.

The second task is to evaluate the classification results of SVM classifiers. The purpose of data publication is to conduct data analysis and data mining. We adopt SVM to evaluate the data utility from the perspective of data applications, as SVM is the most popular classification approach among various data mining techniques with powerful discriminative features both in linear and non-linear classifications [35]. Specifically, we train two classifiers on ACS to predict whether an individual: (1) goes to school, (2) lives in a multi-generation family; four classifiers are constructed on NLTCS to predict whether an individual: (1) is unable to go outside, (2) is unable to manage money, (3) is unable to bathe, and (4) is unable to travel; two classifiers are trained on BR2000 to predict whether an individual (1) owns a private dwelling, (2) is a Catholic; and two classifiers are trained on Adult to predict whether an individual (1) is a female, (2) makes over K a year. For each classifier, we use of the tuples of the dataset for training and the other as the testing set. The prediction accuracy of each SVM classifier is measured by the *misclassification rate* on the testing set.

6.1.3 Comparison Approaches

For the two evaluation metrics mentioned above, we compare our mechanism DP2-Pub with two existing approaches: (1) PrivBayes [9], which first constructs a Bayesian network to model the correlations among the attributes in a dataset, then injects noise into each marginal distribution in the Bayesian network to realize differential privacy, and finally constructs an approximation to the data distribution of the original dataset using the Bayesian network and the noisy marginal distributions; (2) JTree [10], which first develops a robust sampling-based framework to systematically explore the dependencies among all attributes based on the junction tree algorithm and subsequently build a dependency graph; (3) DPPro [12], which projects a -dimensional vector representation of a user's attributes into a lower -dimensional space by a random projection, and then adds noise to each resultant vector. Note that we choose PrivBayes, JTree and DPPro for our comparison study because the first two are benchmark solutions in a way of decomposing high-dimensional data into a set of low-dimensional marginal distributions while the latter is an effective approach of random projection.

6.1.4 Parameter Settings

In our experiments, we use DP-Pub to denote the case with a trusted server and DP-Pub the one with a semi-honest server. The privacy budget of DP-Pub is evenly distributed to the two phases, i.e., – . For DP-Pub , there is no need to partition the privacy budget since the data is first locally differentially privatized, i.e., the privacy budget is completely allocated to the local privacy procedure. For the parameter used in the construction of the Bayesian network, we test . Since the time cost for larger values is typically higher, we do not try the cases of . Based on our experiments, we observe that the influence of on the experimental results is not obvious. The reason possibly lies in that the structure of the Markov blanket can help to accurately learn the data correlations between different attributes. In the following section, we present the experimental results of .

6.2 Experimental Results

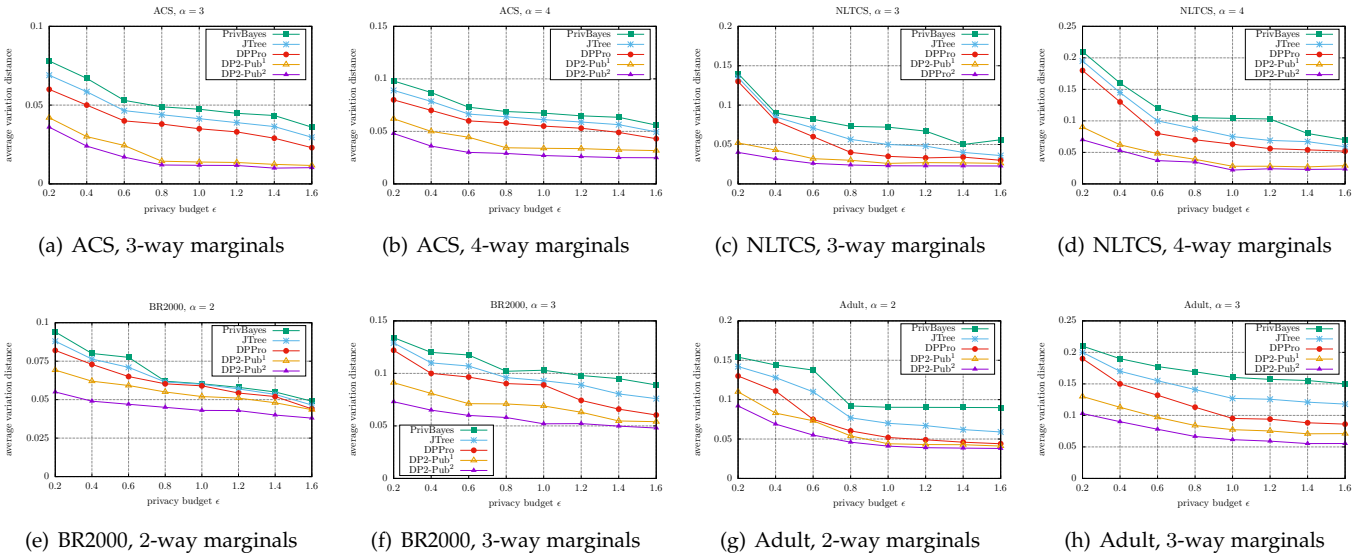
In this subsection, we carry out independent runs for each of the experiments mentioned above and report the averaged results for statistical confidence.

6.2.1 Results on Average Variation Distance

For the task of examining the accuracy of -way marginals, we compute all the -dimensional attribute unions and compare the averaged variation distance of PrivBayes, JTree, DPPro, DP-Pub and DP-Pub , with a varying privacy budget from to .

Figure 5 shows the average results of the variation distance of each approach on the four datasets. From Figure 5, one can see that the average variation distances of these three approaches decrease when increases over the four datasets. It is obvious that when is larger, smaller noise is required, and the data utility is higher. One can also observe that our approach clearly outperforms PrivBayes and DPPro in all cases for ACS and NLTCS, while for BR2000 and Adult, the relative superiority is more pronounced when is small. There are several reasons that DP2-Pub outperforms PrivBayes, JTree and DPPro. First, PrivBayes constructs a Bayesian network while JTree adopts a junction tree algorithm to model the data correlation, and both of them generate a set of noisy conditional distributions of original datasets. That is, for each attribute-parent pair, both PrivBayes and JTree generate differentially private conditional distributions by adding Laplace noise which makes the data utility of the dataset drastically decrease. In our approach, we only utilize the Bayesian network to learn the correlations between different attributes and adopt our proposed invariant post randomization to achieve data perturbation, which ensures that there is almost no loss of statistical information. The probability distribution of each attribute variation is basically unchanged after the double-perturbation. Second, the random projection method DPPro does not consider the data characteristics and only preserves the pairwise distance when generating the random projection matrix, thus it may lead to relatively low utility especially when there exist data correlations between different attributes. In our approach DP2-Pub, we learn the data correlations of the original dataset and consider the importance of different attributes when allocating the privacy budget.

DP-Pub performs better than DP-Pub according to the results shown in Figure 5. This is counter-intuitive as centralized differential privacy usually performs better than local differential privacy because centralized differential privacy adds noise based on the sensitivity of a particular



(a) ACS, 3-way marginals (b) ACS, 4-way marginals (c) NLTCS, 3-way marginals (d) NLTCS, 4-way marginals

(e) BR2000, 2-way marginals (f) BR2000, 3-way marginals (g) Adult, 2-way marginals (h) Adult, 3-way marginals

Fig. 5. Results of α-way marginals with different ϵ.

query function while in local differential privacy noise is added via post randomization. But in DP-Pub , noise is added for differentially private Bayesian network construction and for post randomization, with none of them considering the sensitivity of a particular query function, which is more general at the cost of lower utility. Moreover, at the same budget level, adding noise at two phases increases the total amount of noise as the added noise amount is not linearly proportional to the privacy budget – it is super-linear, which also contributes to the lower utility of DP-Pub .

6.2.2 Results on SVM classification

For the second task, we evaluate the performance of PrivBayes, JTree, DPPro, DP2-Pub DP-Pub , and Non-Private (no DP is considered) for SVM classification. Figure 6 shows the misclassification rate of each approach under different privacy budgets. One can see that the error of Non-Private remains unchanged for all since it does not consider differential privacy. One can also see that both DP-Pub and DP-Pub outperform PrivBayes, JTree and DPPro on almost all datasets. The reason for the higher classification accuracy of our approach lies in that it can achieve higher data utility with a better retention of correlations among attribute variables and a higher accuracy of joint distributions. More specifically, both DP-Pub and DP-Pub retain the data characteristics while satisfying privacy guarantee, thus can help to obtain good results of SVM classifications. Moreover, the misclassification rate decreases faster when increases from to , and the decrease of the misclassification rate is not obvious when is larger than . This indicates that a higher privacy level with a small leads to a lower data utility.

7 Conclusions and Future Research

In this paper, we propose a differentially private data publication mechanism DP2-Pub consisting of two phases, attribute clustering and data randomization. Specifically, in the first phase, we present the procedure of attribute clustering using the Markov blanket model based on the differentially private Bayesian network to achieve attribute clustering and obtain a reasonable allocation of privacy budget. In the second phase, we design a detailed invariant post randomization method by conducting a double-perturbation while satisfying local differential privacy. Our privacy analysis shows that DP2-Pub satisfies differential privacy. We also extend our mechanism making it suitable for the scenario with a semi-honest server in a local-differential privacy manner. Comprehensive experiments on four real-world datasets demonstrate that DP2-Pub outperforms existing methods and improves data utility with strong privacy guarantee.

In our future research, we intend to combine other effective dimensionality reduction techniques [36, 37] with differential privacy to investigate their impact on the data utility of published data. Particularly, we intend to combine DP with manifold learning [36], which is a popular approach for non-linear dimensionality reduction that maps a high dimensional data space into a low-dimensional manifold representation of the data while preserving a certain form of geometric relationships between the data points.

Acknowledgment

This work was partially supported by the US National Science Foundation under grant CNS-1704397.

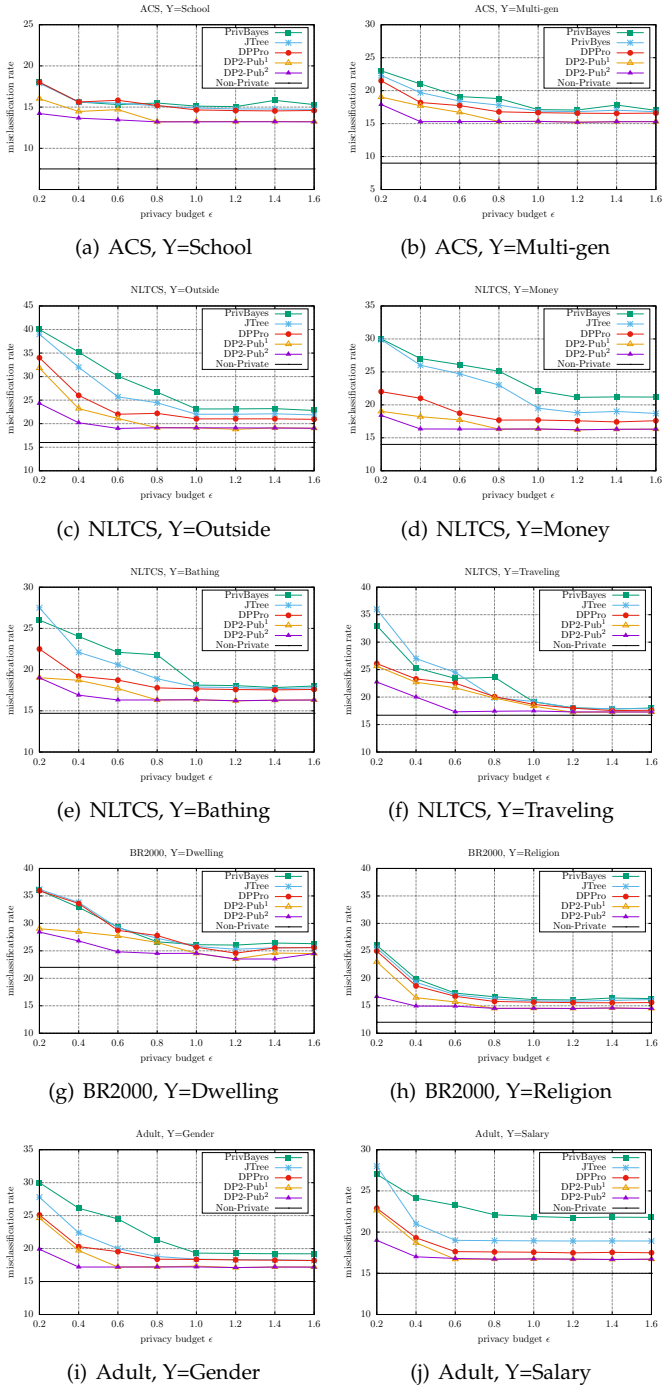
References

[1] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, vol. 101, pp. 417–428, 2019.

[2] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (Csur)*, vol. 42, no. 4, pp. 1–53, 2010.

[3] R. Gay, P. Méaux, and H. Wee, "Predicate encryption for multi-dimensional range queries from lattices," in *IACR International Workshop on Public Key Cryptography*. Springer, 2015, pp. 752–776.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.



(a) ACS, Y=School (b) ACS, Y=Multi-gen

(c) NLTCS, Y=Outside (d) NLTCS, Y=Money

(e) NLTCS, Y=Bathing (f) NLTCS, Y=Traveling

(g) BR2000, Y=Dwelling (h) BR2000, Y=Religion

(i) Adult, Y=Gender (j) Adult, Y=Salary

Fig. 6. Results of SVM with different ϵ

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[6] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[7] D. Cynthia, "Differential privacy," *Automata, languages and programming*, pp. 1–12, 2006.

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

[9] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.

[10] R. Chen, Q. Xiao, Y. Zhang, and J. Xu, "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, pp. 129–138.

[11] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in *Advances in database technology: proceedings. International conference on extending database technology*, vol. 2014. NIH Public Access, 2014, p. 475.

[12] C. Xu, J. Ren, Y. Zhang, Z. Qin, and K. Ren, "Dppro: Differentially private high-dimensional data release via random projection," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 3081–3093, 2017.

[13] X. Cheng, P. Tang, S. Su, R. Chen, Z. Wu, and B. Zhu, "Multi-party high-dimensional data publishing under differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[14] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 729–745.

[15] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pp. 1054–1067.

[16] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 2015, pp. 127–135.

[17] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41–61, 2016.

[18] C. Ju, Q. Gu, G. Wu, and S. Zhang, "Local differential privacy protection of high-dimensional perceptual data by the refined bayes network," *Sensors*, vol. 20, no. 9, p. 2516, 2020.

[19] N. Wang, X. 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 International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 638–649.

[20] J. Domingo-Ferrer and J. Soria-Comas, "Multi-dimensional randomized response," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[21] S. Takagi, T. Takahashi, Y. Cao, and M. Yoshikawa, "P3gm: Private high-dimensional data release via privacy preserving phased generative model," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 169–180.

[22] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[23] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[25] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, pp. 94–103.

[26] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[27] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[28] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[29] T. K. Nayak and S. A. Adeshiyan, "On invariant post-randomization for statistical disclosure control," *International Statistical Review*, vol. 84, no. 1, pp. 26–42, 2016.

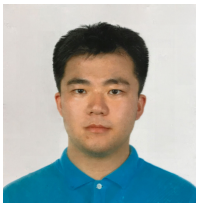
[30] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 2009, pp. 19–30.

[31] M. S. Ali, K. Dolui, and F. Antonelli, "Iot data privacy via blockchains and ipfs," in *Proceedings of the seventh international conference on the internet of things*, 2017, pp. 1–7.

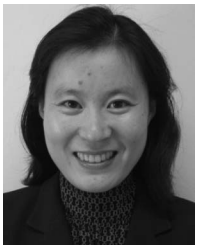
[32] K. G. Manton, "National long-term care survey: 1982, 1984, 1989, 1994, 1999, and 2004," *Inter-university Consortium for Political and Social Research*, 2010.

[33] S. Ruggles, K. Genadek, R. Goeken, J. Grover, and M. Sobek, "Integrated public use microdata series: Version 6.0," *Retrieved from https://international.ipums.org.*, 2015.
[34] K. Bache and M. Lichman, "UCI machine learning repository," 2013.
[35] N. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: a survey," *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 24–38, 2009.
[36] P. Orzechowski, F. Magiera, and J. H. Moore, "Benchmarking manifold learning methods on a large collection of datasets," in *European Conference on Genetic Programming (Part of EvoStar)*. Springer, 2020, pp. 135–150.
[37] R. Li, Y. Shi, Y. Han, Y. Shao, M. Qi, and B. Li, "Active and compact entropy search for high-dimensional bayesian optimization," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
vspace-5 mm

Honglu Jiang received her Ph.D. degree in computer science from The George Washington University in 2021. Currently she is an assistant professor in the Department of Computer Science and Software Engineering at Miami University. Her research interests include wireless and mobile security, privacy preservation, federated learning and data analytics. She is a senior member of IEEE.



Haotian Yu received his B.A. degree from the University of Minnesota in 2017, and M.S. degree in Data Analytics from The George Washington University in 2020. His research interests include data analysis for social networks.



Xiuzhen Cheng received her M.S. and Ph.D. degrees in computer science from the University of Minnesota—Twin Cities in 2000 and 2002, respectively. She is a professor of Computer Science at Shandong University, P. R. China. Her current research focuses on Blockchain computing, privacy-aware computing, and wireless and mobile security. She served/is serving on the editorial boards of several technical journals and the technical program committees of various professional conferences/workshops. She was a faculty member in the Department of Computer Science at The George Washington University from September 2002 to August 2020, and worked as a program director for the US National Science Foundation (NSF) from April to October in 2006 (full time) and from April 2008 to May 2010 (part time). She received the NSF CAREER Award in 2004. She is a member of ACM, and a Fellow of IEEE.

Jian Pei is currently Professor at Duke University. His professional interest is to facilitate efficient, fair, and sustainable usage of data and data analytics for social, commercial and ecological good. Through inventing, implementing and deploying a series of data mining principles and methods, he produced remarkable values to academia and industry. His algorithms have been adopted by industry, open source toolkits and textbooks. His publications have been cited more than 107,000 times. He is also an active and productive volunteer for professional community services, such as chairing ACM SIGKDD, running many premier academic conferences in his areas, and being editor-in-chief or associate editor for the flagship journals in his fields. He is recognized as a fellow of the Royal Society of Canada (i.e., the national academy of Canada), the Canadian Academy of Engineering, ACM, and IEEE. He received a series of prestigious awards, such as the ACM SIGKDD Innovation Award, the ACM SIGKDD Service Award, and the IEEE ICDM Research Award.

Robert Pless received the BS degree in computer science from Cornell University in 1994 and the PhD degree in computer science from the University of Maryland in 2000. He is a Patrick and Donna Martin Professor of Computer Science at The George Washington University. His research interests focus on computer vision with applications in environmental science, medical imaging, robotics and virtual reality. He chaired the IEEE Workshop on Omnidirectional Vision and Camera Networks (OMNIVIS) in 2003, and the MICCAI Workshop on Manifold Learning in Medical Imagery in 2008. He received the NSF career award in 2006.

Jiguo Yu received the Ph.D. degree from the School of Mathematics, Shandong University, in 2004. He became a Full Professor in the School of Computer Science, Qufu Normal University, Shandong, China, in 2007, and currently is a Full Professor at the Qilu University of Technology, Jinan, Shandong China. His main research interests include blockchain, IoT security, privacy-aware computing, wireless distributed computing, and graph theory. He is a Fellow of IEEE, a member of ACM, and a Senior Member of China Computer Federation (CCF).