SCIENCE CHINA
Information Sciences

January 2023, Vol. 66 112101:1–112101:18
https://doi.org/10.1007/s11432-022-3583-x

• RESEARCH PAPER •
From CAS & CAE Members

Locally differentially private high-dimensional data synthesis

Xue CHEN[1,2], Cheng WANG[1,2], Qing YANG[1,2], Teng HU[1,2] & Changjun JIANG[1,2*]

[1]*Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai* 201804, *China;*
[2]*National (Province-Ministry Joint) Collaborative Innovation Center for Financial Network Security, Tongji University, Shanghai* 201804, *China*

Received 14 March 2022/Revised 26 May 2022/Accepted 29 August 2022/Published online 26 December 2022

Abstract In local differential privacy (LDP), a challenging problem is the ability to generate high-dimensional data while efficiently capturing the correlation between attributes in a dataset. Existing solutions for low-dimensional data synthesis, which partition the privacy budget among all attributes, cease to be effective in high-dimensional scenarios due to the large-scale noise and communication cost caused by the high dimension. In fact, the high-dimensional characteristics not only bring challenges but also make it possible to apply some technologies to break this bottleneck. This paper presents SamPrivSyn for high-dimensional data synthesis under LDP, which is composed of a marginal sampling module and a data generation module. The marginal sampling module is used to sample from the original data to obtain two-way marginals. The sampling process is based on mutual information, which is updated iteratively to retain, as much as possible, the correlation between attributes. The data generation module is used to reconstruct the synthetic dataset from the sampled two-way marginals. Furthermore, this study conducted comparison experiments on the real-world datasets to demonstrate the effectiveness and efficiency of the proposed method, with results proving that SamPrivSyn can not only protect privacy but also retain the correlation information between the attributes.

Keywords local differential privacy, high-dimensional data synthesis, sampling, data privacy, privacy-preserving protocols

Citation Chen X, Wang C, Yang Q, et al. Locally differentially private high-dimensional data synthesis. Sci China Inf Sci, 2023, 66(1): 112101, https://doi.org/10.1007/s11432-022-3583-x

1 Introduction

With the development of the Internet and big data technologies, data can be used to capture user behavior and play a very important role in all fields of society [1, 2]. It is becoming more commonplace for data-centric organizations to collect and analyze users' data records to provide better decisions and customized services. For example, in clinical medicine, researchers have utilized clinical and genetic data from base populations to estimate the appropriate warfarin dose [3]. Amazon built effective recommender systems on a huge customer database composed of more than 300 million users to recommend personalized offers, products, and discounts based on their purchasing and browsing histories [4]. Although collecting and analyzing such data records boosts convenience, they can be leaked causing serious consequences (e.g., Facebook[1)], AOL[2)]). To deal with this problem, local differential privacy (LDP) has been proposed and has attracted extensive interest owing to the stringent privacy guarantees it provides [5–8]. In the local model, the privatized process transfers from the server-side to the client-side, which ensures that the original data can only be accessed by the users themselves. Therefore, in recent years, some companies have deployed LDP technology in their products to collect user data to provide better services while protecting users' privacy. For instance, Google [9] was the first to deploy LDP in Chrome to collect

* Corresponding author (email: cjjiang@tongji.edu.cn)
1) https://www.comparitech.com/blog/information-security/267-million-phone-numbers-exposed-online/.
2) https://en.wikipedia.org/wiki/AOL_search_data_leak.

users' browser settings to block malicious software that modifies the configuration without user consent. Apple [10] implemented LDP in its operating systems for discovering popular emojis in different countries to provide customized emoji recommendations for users to improve their experience.

In the academic field, the existing work on implementing LDP has primarily focused on devising tailored algorithms for specific analysis tasks on low-dimensional data, such as frequency estimation [9, 11–14], mean estimation [15], and heavy hitter identification [8,16]. For new tasks, new LDP-enabled algorithms need to be designed; however, this method is not only time-consuming but also has poor scalability. High-dimensional data contain rich knowledge and are gradually playing an important role; thus, there is an urgent need to study them. For example, in the finance field, human characteristics are characterized by multiple dimensions. Therefore, an efficient solution to address the above problems is to synthesize a high-dimensional dataset under the constraint of LDP that captures the correlations between attributions in the original dataset. Consequently, the synthetic dataset can perform any task without privacy concerns or modification to existing algorithms, and the information in high-dimensional data can be fully mined. Unfortunately, the combination of LDP with high-dimensional data synthesis poses significant challenges in both data utility and communication cost. Because they are high dimensional, existing classical methods of generating low-dimensional data, which partition the privacy budget among all attributes, cease to be effective. Specifically, as each attribute shares a given privacy budget, that allocated to each attribute could be quite small, resulting in significant noise being added to the data. In addition, in the high-dimensional scenario, the privatized data record sent to the server leads to a huge communication cost, and in particular, the complex privatization perturbation algorithm aggravates this situation. For example, LoPub [17] was proposed for multidimensional data publication based on RAPPOR [9] and the expectation-maximization (EM) algorithm. It partitions the privacy budget evenly to each attribute of a data record, and then encodes each attribute onto a k-length Bloom filter and applies randomized response (RR) technology [18] to perturb each bit of the Bloom filter. The Bloom filters of all attributes of a data record are spliced together, and finally sent to the server. When the dimension is relatively high, LoPub is inefficient in terms of data utility and communication cost.

Although the high-dimensional characteristic brings significant challenges to data synthesis, it also makes it possible to apply some technologies to break this bottleneck. This study proposes a method called SamPrivSyn for locally differentially private high-dimensional data synthesis via sampling technology. SamPrivSyn is composed of two modules: one is the marginal sampling module, which is used to sample from user data records to obtain two-way marginals. To preserve the correlation between attributes, the sampling criterion is based on mutual information. The higher the mutual information of attribute pairs, the higher the attribute pair sampled. Additionally, the mutual information is constantly updated according to the sampled data, as the local data cannot be obtained in advance under LDP. The second marginal is the data generation module, which is used to construct the synthetic dataset, and is based on the two-way marginals. In SamPrivSyn, instead of using all marginals, only the low-degree marginals are used to represent the high-dimensional dataset, which can effectively reduce both the noise added to each data record and the communication cost. In other words, SamPrivSyn can deal with any high-dimensional dataset, and its noise and communication cost are not greatly increased by the high dimension. Therefore, high data utility can be maintained while providing strong privacy protection at a low communication cost. The data generation algorithm in SamPrivSyn can also efficiently synthesize the dataset based on these marginals while retaining the correlations between attributes.

The overall contributions of this study can be summarized by the following points.

• **Marginal sampling:** An efficient sampling method based on constantly updated mutual information for obtaining two-way marginals from the high-dimensional dataset to represent the original dataset.

• **Data generation:** An iterative and effective data generation method for constructing a synthetic dataset that retains the correlations between attributes.

• **Experimental evaluation:** An extensive experimental regime to verify the effectiveness of SamPrivSyn.

The remainder of this paper is organized as follows. Section 2 describes the recent related work. Section 3 introduces some basic knowledge. Section 4 elaborates the proposed method. Section 5 presents the experimental evaluation. And Section 6 concludes the study.

2 Related work

Differential privacy was first defined by Dwork et al. [5,19]. Traditional differential privacy, also known as centralized differential privacy (CDP) [20–22], depends on the condition that the data collector/server is a trusted third party and will not leak the users' privacy information. However, an absolutely trusted third party is nonexistent in reality. Even if the servers/data collectors claim that they will not leak users' sensitive information, the privacy of users is still not guaranteed, and therefore LDP [6–8] has been put forward to deal with the above issue. Instead of the above assumption, LDP provides stronger privacy guarantees as only the data owner has access to the original data, and the privacy disclosure risk caused by untrusted third parties is eliminated. Numerous LDP-enabled algorithms [6,15,23] have been put forward by various authors. The process of LDP can be summarized as encoding, perturbation, aggregation, and estimation. The current research on LDP mainly focuses on statistical tasks, e.g., frequency estimation, mean estimation, and range query. Using information-theoretic converse techniques, Duchi et al. [6] were the first to provide lower and upper bounds on convex statistical estimation under the LDP setting.

2.1 Frequency estimation

Frequency estimation is a classical statistical tool for analysis and, under LDP, is one of the widely studied topics. Frequency estimation under LDP mainly focuses on discrete data and non-numeric data, and the research hotspot is the design of efficient perturbation mechanisms that satisfy LDP. RR [18] is one of the building blocks for binary data collection under LDP. To deal with multi-valued data, k-ary randomized response (kRR) [24] was proposed, which applies RR technology to arbitrary number of possible values. However, the estimated performance drops sharply with an increase in the dimension. For single-attribute frequency estimation, RAPPOR [9] was proposed by Google to collect user data under LDP. This approach is hashed-based, in which each value is encoded to a Bloom filter, and each bit of the Bloom filter is perturbed with the RR mechanism. Similarly, optimal local hash (OLH) [25] optimizes the choices for hash functions to improve the estimation accuracy, while optimized unary encoding (OUE) [25] typically has better performance when the dimension of the attributes is high. It encodes the original value u_i as a length d binary vector, and only the v-th position is set to 1. The RR mechanism is then applied to perturb each bit of the vector. Significant research has been focused on the joint distribution estimation of multi-attributes. For example, LoPub [17], which is a combination of RAPPOR and EM, was proposed for joint distribution estimation. CALM [26] was proposed for k-marginal estimation by sending m size-l marginals. It uses maximum entropy to estimate the marginals that cannot be directly obtained by adding the m size-l marginals. Additional work has focused on heavy hitter identification, aiming to find the items with frequency over a given threshold. For example, LDPMiner [16] was proposed for the heavy hitter identification over set-valued data, and was a two-phase framework. One is to select a candidate set of top frequent items from an entire dataset, and the other is to identify the exact frequent item from the candidate set. Wang et al. [8] proposed a method of overlapping fragments. Specifically, users in each group report a prefix of their values of a given length, and the data collector then iteratively looks for frequent items.

2.2 Mean estimation

Mean estimation under LDP consists of adding noise to the original value, and then offsetting the noise by aggregating the whole perturbed result such that the statistical results are unbiased estimates. Duchi et al. [6] first proposed a novel method called MeanEst for mean estimation under LDP. This method transforms users' value tuples into binary variable tuples through a specific probability distribution, and then applies RR technology while ensuring that the final statistical result is unbiased. However, MeanEst is only applicable to numeric attributes. To address this limitation, the piecewise mechanism (PM) and hybrid mechanism (HM) [27] were proposed for both numeric and categorical attribute estimation under LDP. In addition, harmony-mean [7] was proposed to simplify Duchi's method and reduce the data transmission cost with sampling technology. The problem is to query the mean under the LDP setting over the key-valued data. To address this issue, Ye et al. [28] generated random values from $[-1, 1]$ and then updated these random values to make the mean estimation closer to the real one. Gu et al. [29] generated random values from $\{-1, 1\}$ with a uniform probability of 0.5, with the expectation of it being 0 and having no impact on the statistical sum. Moreover, they improved the estimation performance by designing an effective privacy budget allocation mechanism. In the following work, Sun

Table 1 Notation description

Notation	Description	Notation	Description
R	Randomized function	d	The number of attributes
n	The total data records	U^i	The ith user's data record
A_i	The ith attribute	u_j^i	The ith user's value on A_j
Ω_i	The domain of A_i	ω_i^j	The jth value of Ω_i
ϵ	Privacy budget	Uniform(a, b)	Uniform distribution

et al. [30] proposed a method to encode continuous numerical values onto extreme values -1 or 1, and then sampled from the values constructed by all the possible combinations between the key and the value.

2.3 Range query

Range query under LDP consists of computing the fraction of a specified interval value, accounting for a population. To limit the variance, the current solutions are based on hierarchy. For instance, Cormode et al. [31] first formalized and studied the range query under the LDP setting by applying a hierarchical approach to construct a binary tree, then utilized postprocessing to improve the accuracy. Wang et al. [32] introduced a hierarchical decomposition scheme for multidimensional range queries by transforming the ordinal dimensions into sub-intervals, which can reduce the worst-case squared error. Du et al. [33] presented a dynamic LDP-enabled range query method called AHEAD, which can adaptively determine the granularity of the domain composition and decrease the impact of insertion noise.

Notably, all of the above methods are aimed at low-dimensional data. However, in reality, high-dimensional data are fast becoming more commonplace and possess research value. The current research ideas are to design corresponding LDP algorithms based on specific tasks, which are inefficient and have poor scalability. Therefore, to address the above limitations, SamPrivSyn aims to synthesize high-dimensional datasets under the constraint of LDP. Compared with data synthesis under differential privacy with that under LDP, LDP-enabled data synthesis is more challenging, as little information about the dataset can be used under the constraint of LDP, and it is impossible to obtain the distribution of the dataset in advance. For example, one of the representative studies on data synthesis under differential privacy is PrivSyn [34], which also consists of two modules, namely marginal selection and data synthesis. In marginal selection, PrivSyn applies a greedy algorithm that selects the pairs to form marginals, and the greedy algorithm needs to obtain the distribution of data in advance to measure the correlation between the attributes. In the data synthesis, PrivSyn applies the graduate update method (GUM) to update the randomly generated dataset by adopting a combination of replication and duplication operations. The replace operation then directly replaces those over-counted records with the under-counted records, and the duplication discards an existing record and substitutes an existing record. In the marginal selection, PrivSyn selects marginals based on the correlation that is obtained by the distribution. However, in the LDP setting, it is impossible to obtain the distribution, let alone the correlation between attributes. Therefore, SamPrivSyn adopts a step-by-step update strategy. For each batch sampled, the correlations of attributes are updated according to the whole sampled record. The probability distribution of sampling is then updated according to the updated correlation to sample the next batch of users with the updated probability, and iterates according to this process until all users are sampled. Compared with PrivSyn, SamPrivSyn synthesizes data more efficiently, as it does not directly replace the over-counted records with the under-counted records. SamPrivSyn first updates the one-way marginal distribution so that the one-way marginal is close to the real one, and then updates the two-way marginal distribution under the condition of keeping the one-way marginal unchanged.

3 Preliminaries

This section introduces some basic foundations, including the definition of LDP and the randomized mechanism. Some notations used in this paper are listed in Table 1. For example, R represents the randomized function, and A_i denotes the i-th attribute.

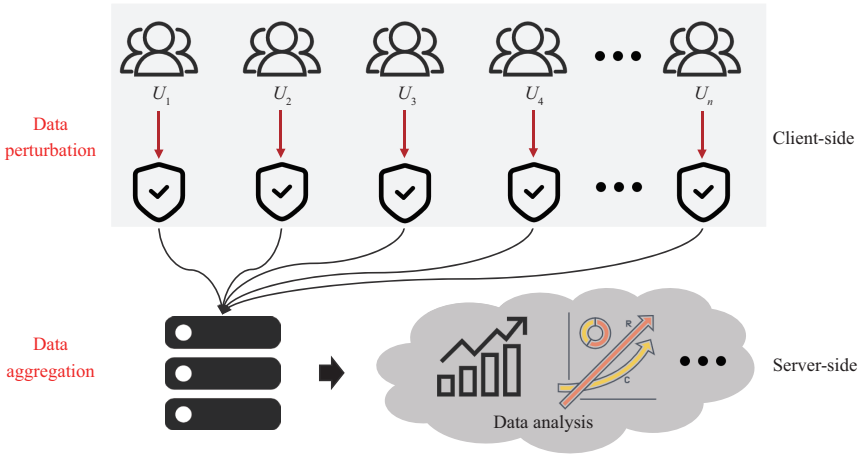

Figure 1 (Color online) Process of locally differentially private data collection.

3.1 Local differential privacy

Figure 1 illustrates the privatized process of the privacy-preserving model under the constraint of LDP. In the LDP-enabled model, each individual's data record is defined as U^i and is perturbed by a randomized function defined as $\mathcal{R}$. The randomized result $\mathcal{R}(U^i)$ is sent to the server/data collector for analysis. Given a privacy budget ϵ that determines the privacy-preserving level, namely a smaller ϵ indicates better privacy protection, the definition of ϵ-LDP can be given.

Definition 1 (ϵ-LDP [6])**.** Given a randomized perturbation function $\mathcal{R}$, it can be said that $\mathcal{R}$ satisfies ϵ-LDP if and only if for any two data records, denoted by U^i and U^j, and any outputs $U^* \in \text{Range}(\mathcal{R})$, it holds that

$$\Pr(\mathcal{R}(U^i) = U^*) \leqslant \mathrm{e}^{\epsilon} \times \Pr(\mathcal{R}(U^j) = U^*). \tag{1}$$

3.2 k-ary randomized response

kRR [24] mechanism is a classical randomized mechanism for perturbing data while satisfying LDP. The definition of kRR is formalized as follows.

Definition 2 (kRR [24])**.** Let U be a k-valued attribute, where each user takes a value $u \in U$ on this attribute. U is stochastically mapped onto Y (i.e., $Y = U$) and the perturbed value $y \in Y$ of the user is then sent to the data collector. Then, for any u and y,

$$\Pr(y|u) = \begin{cases} \frac{\mathrm{e}^{\epsilon}}{k-1+\mathrm{e}^{\epsilon}}, & y = u, \\ \frac{1}{k-1+\mathrm{e}^{\epsilon}}, & y \neq u. \end{cases} \tag{2}$$

In the kRR mechanism, k is the domain size of the attribute, where each user sends their original value, i.e., the true value, with the probability $\frac{\mathrm{e}^{\epsilon}}{k-1+\mathrm{e}^{\epsilon}}$ and then sends other values (the remaining values in the domain) with $\frac{1}{k-1+\mathrm{e}^{\epsilon}}$.

4 Method

This section further elaborates on the proposed method SamPrivSyn for high-dimensional data synthesis under the constraint of LDP via sampling technology. To generate high-dimensional data in a locally differentially private way, the problem remains how to reduce the excessive noise and communication cost caused by high-dimensional characteristics. One promising approach is to use low-degree joint distributions, namely, marginals, to represent the high-dimensional dataset. With this concept in mind, SamPrivSyn is composed of two modules, the first is the marginal sampling module, which is executed on the client-side to collect two-way marginals to represent the original dataset; the second is the data generation module that is acted on the server-side to synthesize the dataset. Figure 2 illustrates the framework of SamPrivSyn. In the marginal sampling module, the sampling is based on the continuously updated mutual information.

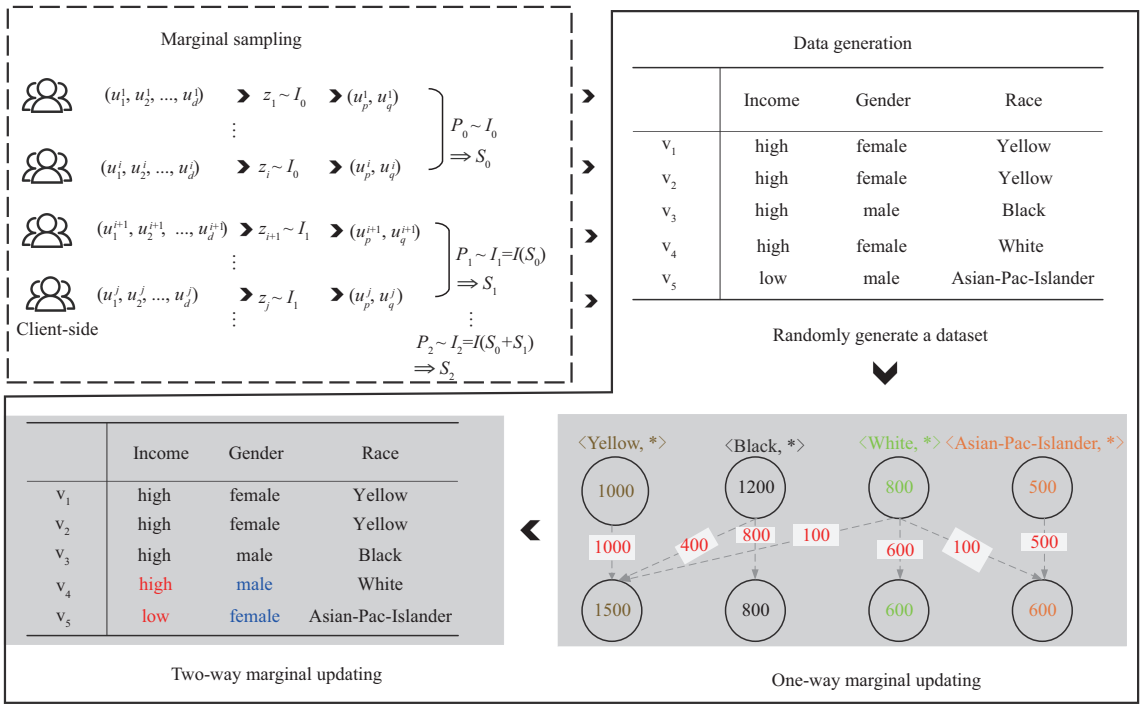

Figure 2 (Color online) Framework of SamPrivSyn, which is composed of two modules. The first is the marginal sampling module for collecting two-way marginals, and the second is the data generation module for synthesizing the dataset.

4.1 Locally differentially private marginal sampling

Therefore, given all the joint distributions of the attributes (i.e., one-way marginals, two-way marginals, ..., d-way marginals), one can synthesize a dataset that is similar to the real dataset. However, in the high-dimensional scenario, computing, or even storing, the full distribution requires exponentially large space. To address this issue, this study proposes an efficient method that uses two-way marginals to represent the dataset. In this way, the correlations between attributes can be greatly retained, as only two attributes share the given privacy budget, and the problem of the communication cost can be relieved.

Given d attributes $(A_1, A_2, \ldots, A_d)$, each user takes a value on each attribute denoted as $U^i = \{u_1^i, u_2^i, \ldots, u_d^i\}$. Instead of sending the complete record to the server, each user only needs to select a pair of attributes. As shown in Algorithm 1, firstly, all the pairs of attributes $L \Leftarrow \{(A_1, A_2), (A_1, A_3), \ldots, (A_{d-1}, A_d)\}$ are first listed as the index of sampling; notably, the attribute pair (A_i, A_j) being the same as (A_j, A_i) is considered, meaning that the number of all attribute pairs can be computed as $m = \text{len}(L) = \frac{d(d-1)}{2}$. To retain the correlations between attributes as much as possible, the attribute pairs are sampled based on mutual information. Specifically, the mutual information of all pairs is first initialized as being equal to $I_i = \frac{2}{d(d-1)}$, that is,

$$I = \left\{ \frac{d(d-1)}{2} \right\}^{\frac{2}{d(d-1)}};$$

then we normalize the mutual information and sample according to the normalized value, e.g.,

$$z_i \sim \left[\frac{I_1}{\sum_{i=1,\ldots,\frac{d(d-2)}{2}} I_i}, \frac{I_2}{\sum_{i=1,\ldots,\frac{d(d-2)}{2}} I_i}, \ldots, \frac{I_{\frac{d(d-2)}{2}}}{\sum_{i=1,\ldots,\frac{d(d-2)}{2}} I_i} \right],$$

where z_i represents the position of the selected attribute pair in L. For example, if $z_i = 1$, it means that the selected pair attribute is $L(z_i) = (A_1, A_2)$. After obtaining the sampling pair $L(z_i)$, the corresponding value $\{u_p^i, u_q^i\}$ of the attribute pair is sampled from the data record U^i. The next step is to add noise to the sampling values to satisfy LDP. As the complex privatized algorithm will bring extra communication cost, in SamPrivSyn, the kRR mechanism is applied to perturb the sampling values. More specifically, given the privacy budget ϵ, for each value of $\{u_p^i, u_q^i\}$, $y_p^i \in Y_p$ is defined as the perturbed version of u_p^i

Algorithm 1 Marginal sampling

Require: n: the number of data records; U_i: the i-th data record; d: the number of attributes; Ω_i: the domain of A_i; ϵ: privacy budget; B: the sampling batch size;
Ensure: S: sampling pairs;
List all pairs of attributions: $L \Leftarrow \{(A_1, A_2), (A_1, A_3), \ldots, (A_{d-1}, A_d)\}$;
The number of all attribute pairs is $m \Leftarrow \frac{d(d-2)}{2}$;
Sampling pairs: $S \Leftarrow [\{\emptyset\}]^m$;
Initialize the mutual information of all pairs:

$$I = \left\{\frac{2}{d(d-2)}\right\}^{\frac{d(d-2)}{2}};$$

for $i = 0, \ldots, n$ **do**
  if $(i+1) \mod B == 0$ **then**
    Update the mutual information based on the sampled data: $I(A_p, A_q) = I(S(A_p, A_q))$;
  else

$$z_i \sim \left[\frac{I_1}{\sum I}, \frac{I_2}{\sum I}, \ldots, \frac{I_{\frac{d(d-2)}{2}}}{\sum I}\right];$$

    Select the values of the pair attributes $L[z_i]$ in U_i: $U_i \sim \{u_p^i, u_q^i\} \Leftarrow (L[z_i] = (A_p, A_q))$;
    Perturb each selected value:

$$Q(y_p^i|u_p^i) = \begin{cases} \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}}, & y_p^i = u_p^i, \\ \frac{1}{|\Omega_p|-1+e^{\epsilon/2}}, & y_p^i \neq u_p^i, \end{cases}$$

$$Q(y_q^i|u_q^i) = \begin{cases} \frac{e^{\epsilon/2}}{|\Omega_q|-1+e^{\epsilon/2}}, & y_q^i = u_q^i, \\ \frac{1}{|\Omega_q|-1+e^{\epsilon/2}}, & y_q^i \neq u_q^i; \end{cases}$$

    Add the perturbed values into S: $S(A_p, A_q) = S[z_i] \cup \{y_p^i, y_q^i\}$;
  end if
end for
Ensure: S.

and the domain of Y_p is equal to Ω_p. The output of u_p^i satisfies the following formula:

$$Q(y_p^i|u_p^i) = \begin{cases} \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}}, & y_p^i = u_p^i, \\ \frac{1}{|\Omega_p|-1+e^{\epsilon/2}}, & y_p^i \neq u_p^i. \end{cases} \tag{3}$$

Eq. (3) means that the privacy budget ϵ is evenly allocated to the two sampling values, and each sampling value flips to any other value of Ω_p with probability $\frac{1}{|\Omega_p|-1+e^{\epsilon/2}}$, and remains unchanged with probability $\frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}}$. Next, the client adds the perturbed sampling values $\{y_p^i, y_q^i\}$ into $S(A_p, A_q)$. After the number of records sampled each time is up to B, which is the given batch size to control the speed of updating the mutual information, the mutual information is updated according to the sampled records S. That is, a LASSO regression is used to estimate the joint distribution of each pair $P(A_p, A_q)$. The estimated process is the same as that in the following Subsection 4.2; it is not repeated. The estimated distribution $\hat{P}(A_p, A_q)$ is then used to compute the mutual information; that is, given the estimated distribution $\hat{P}(A_1, A_2)$,

$$I_1 = \sum_{A_1}\sum_{A_2} \hat{P}(A_1, A_2) \log \frac{\hat{P}(A_1, A_2)}{\hat{P}(A_1)\hat{P}(A_2)}.$$

Next, the updated mutual information is normalized and the B records are sampled accordingly, with the above process repeated until all records are sampled. Finally, the whole sampled record is sent to the server.

Theorem 1. The marginal sampling process of SamPrivSyn satisfies ϵ-LDP.

Proof. Given two arbitrary values $(w_{p_1}, w_{q_1}), (w_{p_2}, w_{q_2})$, and for arbitrary output (y_p, y_q), the ratio of probability that the value $(w_{p_1}, w_{q_1}), (w_{p_2}, w_{q_2})$ outputs the same value (y_p, y_q) after being perturbed by

the kRR mechanism is

$$\begin{aligned}\frac{Q(y_py_q|w_{p_1}w_{q_1})}{Q(y_py_q|w_{p_2}w_{q_2})} &\leqslant \frac{\frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{e^{\epsilon/2}}{|\Omega_q|-1+e^{\epsilon/2}}}{\frac{1}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}}}\\ &= e^{\epsilon/2} \times e^{\epsilon/2}\\ &= e^{\epsilon}.\end{aligned}$$

Therefore, the marginal sampling process of SamPrivSyn satisfies the definition of ϵ-LDP.

4.2 Data generation

After aggregating all perturbed samples S from the client, the server needs to estimate the true two-way marginals, denoted by $\hat{P}(A_i, A_j)$, according to the perturbed samples S. Notably, $S(A_i, A_j)$ represents the set of perturbed values of the attribute pair (A_i, A_j) from all users. Specifically, the server firstly generates the conditional probability matrix $\boldsymbol{Q}_{pq}$ for each attribute pair according to (3). For the attribute pair (A_p, A_q), the conditional probability $Q(y_py_q|\omega_p\omega_q) = Q(y_p|\omega_p) \times Q(y_q|\omega_q)$ is the probability that the true value $\{\omega_p, \omega_q\}$ is perturbed to $\{y_p, y_q\}$. Therefore, $\boldsymbol{Q}_{pq}$ is composed of the conditional probability of each value in combination $\Omega_p \times \Omega_q$. The value of $\boldsymbol{Q}_{pq}$ is given as

$$\boldsymbol{Q}_{pq} = \begin{array}{c} \\ \omega_p^1\omega_q^1 \\ \omega_p^1\omega_q^2 \\ \vdots \\ \omega_p^{|\Omega_p|}\omega_q^{|\Omega_q|} \end{array} \begin{array}{c} \begin{array}{cccc} \omega_p^1\omega_q^1 & \omega_p^1\omega_q^2 & \cdots & \omega_p^{|\Omega_p|}\omega_q^{|\Omega_q|} \end{array} \\ \begin{pmatrix} \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{e^{\epsilon/2}}{|\Omega_q|-1+e^{\epsilon/2}} & \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} & \cdots & \frac{1}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} \\ \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} & \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{e^{\epsilon/2}}{|\Omega_q|-1+e^{\epsilon/2}} & \cdots & \frac{1}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} & \frac{1}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{1}{|\Omega_q|-1+e^{\epsilon/2}} & \cdots & \frac{e^{\epsilon/2}}{|\Omega_p|-1+e^{\epsilon/2}} \times \frac{e^{\epsilon/2}}{|\Omega_q|-1+e^{\epsilon/2}} \end{pmatrix} \end{array}. \tag{4}$$

It can be observed that the distribution of the perturbed marginals (A_p, A_q) is defined as

$$\boldsymbol{M}(A_p, A_q) = \frac{\text{Count}(\omega_p, \omega_q)}{\sum_{(\omega_p,\omega_q)\in(\Omega_p\times\Omega_q)} \text{Count}(\omega_p, \omega_q)}, \tag{5}$$

where $\text{Count}(\omega_p, \omega_q)$ represents the number of (ω_p, ω_q) in all perturbed records $S(A_p, A_q)$. Therefore, the following formula can be obtained:

$$\boldsymbol{M}(A_p, A_q) = P(A_p, A_q)\boldsymbol{Q}_{pq}. \tag{6}$$

The estimated two-way marginal $\hat{P}(A_p, A_q)$ can be computed as

$$\hat{P}(A_p, A_q) = \boldsymbol{M}(A_p, A_q)\boldsymbol{Q}_{pq}^{-1}. \tag{7}$$

A LASSO regression Lasso(M,Q) can therefore be fitted to estimate the distribution of $P(A_p, A_q)$; that is, $\hat{P}(A_p, A_q) = \text{Lasso}(\boldsymbol{M}(A_p, A_q), \boldsymbol{Q}_{pq}^{-1})$. All two-way marginals are obtained in this manner.

Notably, the estimated distribution of SamPrivSyn is an unbiased estimator, and the variance of A_i can be computed. Suppose that the domain of A is $\Omega = \{0, 1\}$. Let P_A be the true probability of 1 in the population U, and p is the probability for the user U^i to respond to the true value, that is, if $U_i = 1$, it outputs 1 with probability p and outputs 0 with probability $1 - p$, and vice versa. Therefore, the following theorem can be obtained.

Theorem 2. The variance of the estimated $\hat{P}_A$ by kRR is

$$\text{var}(\hat{P}_A) = \frac{(pP_A + (1-p)(1-P_A))(p(1-P_A) + (1-p)P_A)}{n(2p-1)^2}.$$

Proof. The perturbed percentage of users who output 1 is

$$\lambda = P(A = 1) = pP_A + (1-p)(1-P_A).$$

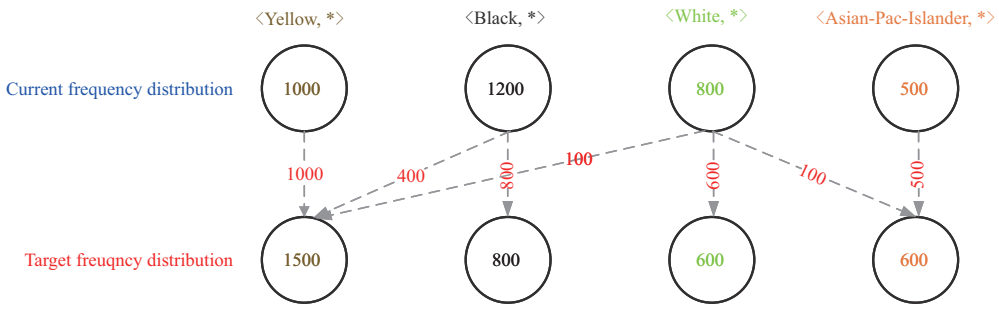

Figure 3 (Color online) Example of one-way updating. The top side is the current frequency distribution of each attribute in the random dataset D_s, and the bottom side is the target frequency distribution, which is calculated from the estimated two-way marginals.

Then, the estimated $\hat{P}_A$ can be computed via

$$\hat{P}_A = \frac{\lambda - (1-p)}{2p-1}.$$

$\hat{P}_A$ is an unbiased estimator, and the variance is

$$\begin{aligned} \mathrm{var}(\hat{P}_A) &= \mathrm{var}\left(\frac{\lambda - (1-p)}{2p-1}\right) && (8)\\ &= \frac{n \times \mathrm{var}(U^i)}{(2p-1)^2 n^2} && (9)\\ &= \frac{(pP_A + (1-p)(1-P_A))(p(1-P_A) + P_A(1-p))}{n(2p-1)^2}. && (10) \end{aligned}$$

The next step is to generate the high-dimensional dataset according to these sampled estimated two-way marginals. Instead of sampling the dataset using marginals, a random dataset D_s is first initialized, and D_s is then updated to ensure that the marginals of D_s are consistent with the estimated marginals as much as possible. The process of updating D_s is divided into two processes; one is to use the one-way marginals to update, and the other is to use the two-way marginals.

4.2.1 *One-way marginal updating*

Firstly, the distribution of each attribute must be computed, namely, one-way marginals, according to two-way marginals.

$$\hat{P}(A_i) = \frac{I(A_i, A_j)}{\sum_{j=1, j\neq i}^{d} I(A_i, A_j)} \hat{P}(A_i, A_j). \quad (11)$$

Then the one-way marginals are transformed to a frequency distribution to update D_s. In particular, given a marginal frequency distribution of an attribute A_i, let $\hat{P}(A_i)$ denote the target estimated distribution, and $C(A_i)$ denote the current random frequency distribution of this attribute in the dataset D_s, and then a graph can be generated. The top side is the current frequency distribution $C(A_i)$ specified by D_s, and the bottom side is the target estimated frequency distribution $\hat{P}(A_i)$ specified by the one-way marginal. Each node represents a value in this attribute associated with the frequency. The current node in D_s is updated according to the target node. For every instance, one attribute is updated and all attributes are iteratively updated.

Figure 3 shows an example of one-way updating. In Figure 3, the attribute is skin color, which has four values {Yellow, Black, White, Asian-Pac-Islander}, the top side is the current frequency distribution of the attribute in D_s, the bottom side is the target estimated distribution (one-way marginal) obtained by (11). Each node represents one value in {Yellow, Black, White, Asian-Pac-Islander} associated with its frequency. The updating process is moving 400 from Black to Yellow, 100 from White to Yellow, and 100 from White to Asian-Pac-Islander. Notably, in one-way updating, only one attribute is updated while all other attributes are kept the same. After updating all attributes, D_s has the similar one-way marginal with the original dataset. In the next step, D_s is updated using two-way marginals, a process called two-way marginal updating.

	Income	Gender	Age
v_1	high	female	adult
v_2	high	female	adult
v_3	high	male	children
v_4	high	female	children
v_5	low	male	children

(a)

	C(*I*,*G*)	*P*(*I*,*G*)
⟨high,female,*⟩	3	2
⟨high,male,*⟩	1	2
⟨low,female,*⟩	0	1
⟨low,male,*⟩	1	0

(b)

	Income	Gender	Age
v_1	high	female	adult
v_2	high	female	adult
v_3	high	male	children
v_4	high	female	children
v_5	low	male	children

(c)

Figure 4 (Color online) Two-way marginal updating. (a) The dataset after the one-way marginal update begins for the two-way update; (b) the two-way marginal distribution of the dataset, marginal table for {Income, Gender}, where the red and blue stand for over-counted and under-counted values, respectively; and (c) the dataset after two-way marginal updating.

4.2.2 *Two-way marginal updating*

This process modifies the two-way marginals of D_s but keeps the one-way marginal the same. Firstly, all the two-way marginals $\hat{P}(A_i, A_j)$ are transformed into a corresponding frequency distribution, denoted by the target frequency. Then, we calculate all two-way marginal distributions in the dataset D_s, where let $C(A_i, A_j)$ denote the two-way marginal distribution of the attribute pair (A_i, A_j). The dataset D_s is then updated according to the difference between the target marginals $\hat{P}(A_i, A_j)$ and the current marginals $C(A_i, A_j)$.

Specifically, only one attribute is updated at a time according to the two-way marginal distribution, and the update of attributes follows their order. At first, the attribute A_1 is fixed and is not modified; then the attribute A_2 is changed to make the distribution of $C(A_1, A_2)$ consistent with that of $\hat{P}(A_1, A_2)$. Analogously, the next attribute A_2 is fixed and the attribute A_3 is changed. The two-way marginal updating process is different from the one-way update, which moves the value with more frequency to the value with less frequency, in that it needs to update the attribute under the condition that the one-way marginal distribution remains unchanged. That is to say, given the target two-way frequency distribution $\hat{P}(A_i, A_j)$ and the current two-way frequency distribution $C(A_i, A_j)$, the attribute A_j is changed to make $C(A_i, A_j)$ similar to $\hat{P}(A_i, A_j)$, but in doing so, keep $A_i, C(A_i), C(A_j)$ unchanged. Therefore, for a value $(\omega_i^p, \omega_j^q) \in (\Omega_i, \Omega_j)$, if $C(\omega_i^p, \omega_j^q) > \hat{P}(\omega_i^p, \omega_j^q)$, namely, the frequency of (ω_i^p, ω_j^q) of D_s is greater than the target frequency, and $C(\omega_i^{p1}, \omega_j^{q1}) > \hat{P}(\omega_i^{p1}, \omega_j^{q1})$, $C(\omega_i^p, \omega_j^{q1}) < \hat{P}(\omega_i^p, \omega_j^{q1})$, $C(\omega_i^{p1}, \omega_j^q) < \hat{P}(\omega_i^{p1}, \omega_j^q)$, then the minimum frequency difference of these frequencies can be calculated; that is,

$$\begin{aligned} \min_{\text{fre}} = \min\{&(C(\omega_i^p, \omega_j^q) - \hat{P}(\omega_i^p, \omega_j^q)), (C(\omega_i^{p1}, \omega_j^{q1}) - \hat{P}(\omega_i^{p1}\omega_j^{q1})), \\ &(\hat{P}(\omega_i^p, \omega_j^{q1}) - C(\omega_i^p, \omega_j^{q1})), (\hat{P}(\omega_i^{p1}, \omega_j^q) - C(\omega_i^{p1}, \omega_j^q))\}. \end{aligned} \tag{12}$$

After meeting all the above criteria, $\min_{\text{fre}}$ records of ω_j^q are changed to ω_j^{q1}. As shown in Figure 4(a), the red represents the record to be changed. In Figure 4(b), the red represents the over-counted values and the blue represents the under-counted values. Specifically, the current frequencies of (high, female) and (low, male) are greater than the target frequencies, and the current frequencies of (high, male) and (low, female) are less than the target frequencies. According to (12), the positions of female in v_4 and male in v_5 need to be exchanged in Figure 4(a), and Figure 4(c) displays the updated result. This method not only ensures the two-way marginal distribution, but also ensures the one-way marginal distribution, so as to make the synthetic dataset capture the correlations in the original dataset as much as possible. When all the joint distributions $P(A_1, A_2), P(A_2, A_3), \ldots, P(A_{d-1}, A_d)$ are updated, the random dataset is updated using the remaining marginals. However, when one attribute is updated, the distributions that are related to this attribute need to be kept unchanged. The update procedure terminates when all

the attributes have been updated. Algorithm 2 shows the complete procedure of data generation.

Algorithm 2 Data generation

Require: S: sampling pairs; Ω_i: the domain of A_i; d: the number of attributes; Ω_i: the domain of A_i; ϵ: privacy budget;
Ensure: D_s: synthetic dataset;
for $i = 0, \ldots, d$ **do**
  for $j = i + 1, \ldots, d$ **do**
    Estimate all the two-way marginal distributions: $\hat{P}(A_i, A_j) = \text{Lasso}(M, Q)$;
  end for
end for
for $i = 0, \ldots, d$ **do**
  Compute the one-way marginal distribution:

$$\hat{P}(A_i) = \frac{I(A_i, A_j)}{\sum_{j=1, j\neq i}^{d} I(A_i, A_j)} \hat{P}(A_i, A_j);$$

end for
for $i = 0, \ldots, d$ **do**
  Update the one-way marginal distribution of D_s: $D_s \Leftarrow C(A_i) - \hat{P}(A_i)$;
end for
for $i = 0, \ldots, d$ **do**
  for $j = i + 1, \ldots, d$ **do**
    Update the two-way marginal distribution of D_s:

$$\begin{aligned} D_s \Leftarrow \min_{\text{fre}} = \min\{&(C(\omega_i^p, \omega_j^q) - \hat{P}(\omega_i^p, \omega_j^q)), (C(\omega_i^{p1}, \omega_j^{q1}) - \hat{P}(\omega_i^{p1}\omega_j^{q1})), \\ &(\hat{P}(\omega_i^p, \omega_j^{q1}) - C(\omega_i^p, \omega_j^{q1})), (\hat{P}(\omega_i^{p1}, \omega_j^q) - C(\omega_i^{p1}, \omega_j^q))\}; \end{aligned}$$

  end for
end for
Ensure: D_s.

5 Experiments

Extensive comparison experiments on a real-world dataset have been performed to evaluate the efficiency and effectiveness of SamPrivSyn. SamPrivSyn was compared with the representative and state-of-art methods, CALM, LoPub, and kRR. The performance was evaluated from two perspectives, namely, the data utility and the communication cost. The data utility is measured by the distribution estimation accuracy and the performance of logistic regression on the synthetic dataset. The distribution estimation accuracy represents the similarity between the estimation joint distribution, while the logistic regression accuracy represents similarity between the training accuracy on the synthetic dataset and the original dataset. Specifically, the distribution estimation accuracy is measured by the commonly-used metric of relative error (RE), which is defined as

$$\text{RE} = \frac{1}{N} \sum \left| \frac{P(A_1 A_2, \ldots, A_d) - \hat{P}(A_1 A_2, \ldots, A_d)}{P(A_1 A_2, \ldots, A_d)} \right|, \tag{13}$$

where N is the size of $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_d$. The performance of the logistic regression is measured by the accuracy, which indicates the proportion of samples that are correctly classified from the total number of samples. Accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{14}$$

where TP, FP, TN, FN represent true positive, false positive, true negative, and false negative, respectively. Notably, the high dimension used in this study is not only related to the dimension of attributes, that is, the number of attributes, but also to the domain size of attributes. The high dimension actually refers to the Cartesian product of the domain size of all attributes, i.e., $|\Omega| = |\Omega_1| \times |\Omega_2| \times \cdots \times |\Omega_d|$. The communication cost is measured by the bits sent to the server-side from the client. For the sake of thoroughness, several groups of experiments were performed to verify the impact of different parameters on performance. More specifically, the impact of privacy budget, data amount, and dimension on the performance of these methods was explored. The experiments were performed on the real-world datasets

Table 2 Dataset description

	Records (n)	Attribute dimension (d)	Domain size ($\|\Omega\|$)
Adult	30128	9	2^{29}
SYN1-Adult	60256	9	2^{26}
SYN2-Adult	120512	9	2^{26}
SYN3-Adult	30128	15	2^{44}
Bank Marketing	45210	17	2^{29}

Adult[3)] and Bank Marketing [4)]. Adult contains personal information, including age, gender, race, salary, and Bank Marketing contains attributes that include age, job, marital status. To evaluate the impact of the different data records and dimensions on the estimation accuracy, the control variable method was adopted to ensure a fair evaluation. Specifically, when evaluating the impact of dimensions on performance, the number of records must be kept the same. On the contrary, when evaluating the impact of the number of records on performance, the dimension must be kept the same. However, in reality, it is difficult to find datasets that meet the above requirements. Therefore, this study extended the existing Adult dataset to several datasets that meet the above requirements to verify the impact of different features on the estimation accuracy. It is worth noting that this study only considered the difference between the distribution of estimation results and the original one, that is, the difference of each value's frequency; therefore, expansion is feasible. Firstly, to study the impact of the number of data records on the performance, the dataset SYN1-Adult was generated by doubling the amount of data for each attribute in the original Adult dataset. Similarly, SYN2-Adult was formed by tripling the amount of data for each attribute in Adult. For exploring the effect of the dimension, six attributes were selected in the original Adult dataset and were added as new attributes to Adult to form a new dataset, SYN3-Adult, which has 15 attributes. Table 2 gives a description of the dataset.

In addition, for a fair comparison, when implementing the comparison methods, the same privacy budget for each method was set. The privacy budget was changed from 1 to 10, and for the parameters of LoPub, the hash function number was set to 2, and the Bloom filter length was set to 64. Subsections 5.1 and 5.2 elaborate on the experimental results from the aspects of data utility and communication cost. Data utility, which is the estimation and logistic regression accuracy, is used to evaluate the effectiveness of the proposed method, and communication cost is used to evaluate the efficiency of SamPrivSyn.

5.1 Data utility

Distribution estimation. Figure 5 illustrates the experimental results on the given datasets. The proposed method was compared with the existing state-of-art LDP-enabled data synthesized methods CALM, LoPub, and kRR. Here, a larger batch size and smaller batch size were set, respectively, but for a fair comparison, the batch sizes for the datasets Adult, SYN1-Adult, SYN2-Adult and SYN3-Adult need to be kept the same. Therefore, the batch size was set to 30000 for Adult, SYN1-Adult, SYN2-Adult, and SYN3-Adult, and 2000 for the dataset Bank Marketing. Since the batch size is a key parameter for the performance, a random value was set for the experiment, and its impact is discussed in a separate section. This study first evaluated the estimation accuracy of the different methods, as shown in Figure 5. The red line, blue line, green line, and the orange line represent CALM, kRR, LoPub, and SamPrivSyn, respectively. In particular, both LoPub and kRR need to build a conditional matrix, which is the probability that the real record changes to others, to estimate the real distribution. Therefore, the size of the conditional matrix is $(|\Omega_1| \times |\Omega_2| \times \cdots \times |\Omega_d|, |\Omega_1| \times |\Omega_2| \times \cdots \times |\Omega_d|)$, which grows with the dimension of the attributes and the domain of each attribute. As a result, the size is so large that the program cannot allocate the amount of memory for it. Hence, the number of combinations of attributes was cut according to the frequency of each record. In CALM, $\binom{2}{d}$ two-way marginals were sent to construct the synthetic dataset. Figures 5(a)–(c) show the comparison results over the datasets with different data amounts. These experiments evaluated the impact of data amount on performance by changing the number of records. The results are shown as follows. Figure 5(a) shows the Adult dataset, whose dimension is 9 and the number of records is 30128; Figure 5(b) shows the SYN1-Adult dataset, whose dimension is 9 and the number of records is 60256; Figure 5(c) shows the SYN2-Adult dataset,

3) Adult dataset. https://archive.ics.uci.edu/ml/datasets/Adult.
4) Bank Marketing. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

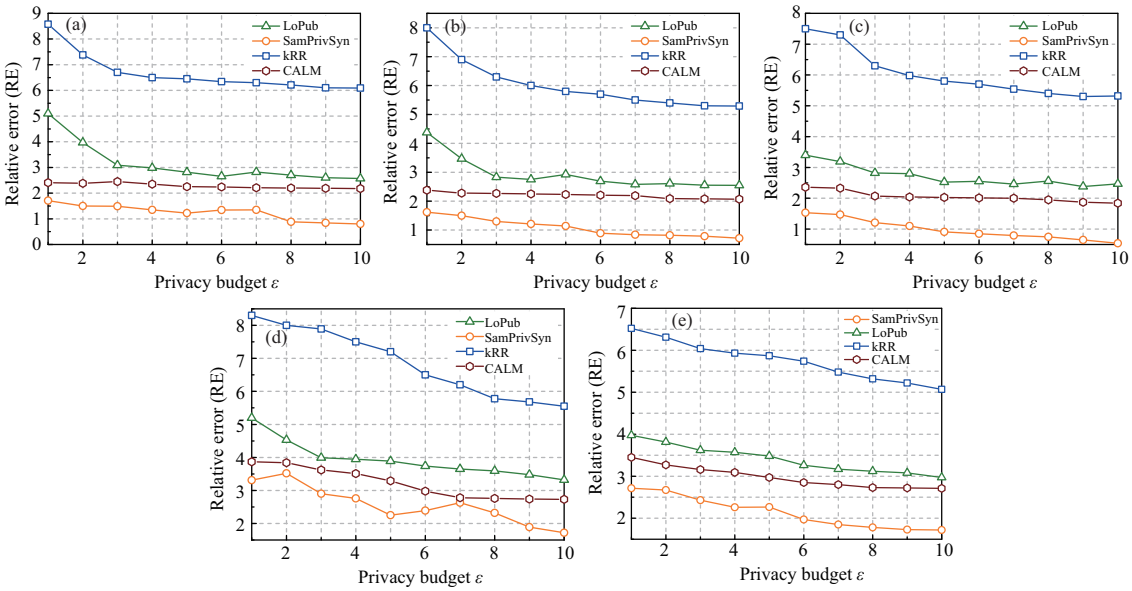

Figure 5 (Color online) Comparison of results for different LDP-enabled data synthesis methods. (a) Adult; (b) SYN1-Adult; (c) SYN2-Adult; (d) SYN3-Adult; (e) Bank Marketing.

whose dimension is 9 and the number of records is 120512; and Figure 5(c) shows the Bank Marketing dataset, whose dimension is 17 and the number of records is 45210.

It can be observed that SamPrivSyn outperformed CALM, LoPub, and kRR, regardless of the number of data records and dimensions, and the estimation accuracy of all methods increased when the privacy budget was increased. Specifically, the performance of kRR was the worst, followed by LoPub, with both of these methods adopting a way of evenly allocating the privacy budget; therefore, when the dimension was relatively high, the privacy budget allocated to each attribute was very small, resulting in more noise added to the dataset, and achieving poor performance. Comparing kRR with LoPub, the performance of kRR was worse because the performance was limited by the domain size Ω_i of the attribute; when the domain size of attributes is high, the probability of kRR remaining unchanged will be very low; that is, the noise added to the dataset will be greater. Moreover, comparing SamPrivSyn with CALM, they both adopted selecting marginals to the data collector; however, CALM applied a uniform selection mechanism; that is, the sampled marginal was selected with the same probability, while SamPrivSyn selected marginals based on mutual information. Therefore, SamPrivSyn was more prominent in preserving the correlation between attributes. In addition, for the distributions that cannot be directly obtained by summing the sampled marginals, CALM applied maximum entropy approach to estimate the joint distribution. However, the maximum entropy takes the estimated marginals as the constraint conditions. Due to the error of the estimated marginals, the constraint conditions may be mutually exclusive, which will affect the accuracy of the final estimation result.

To observe the impact of the number of data records on the estimation accuracy, extensive experiments were conducted with the different data records, and the results are shown in Figures 5(a)–(c). With an increase of the data records, the estimation error of all methods drops, which indicates that the estimation accuracy is positively related to the number of the data records; i.e., the larger the number, the better the performance. This trend is consistent with the law of large numbers. Moreover, the estimation accuracy of CALM is the most stable and will not fluctuate greatly due to the privacy budget or the number of the data records. This is because the distribution estimation method adopted is the nonlinear constrained optimization of maximum entropy, so that the optimization result of maximum entropy will not fluctuate much.

In addition, to explore the impact of dimension on the estimation accuracy, we also conducted experiments, as shown in Figures 5(a) and (d). Comparing Figure 5(a) with Figure 5(d), we can see that the performance of all methods decreases as the dimension increases; the reason is that, for LoPub and kRR, each attribute divides the given privacy budget equally; therefore, when the dimension increases, the privacy budget allocated to each attribute decreases, and then the estimation accuracy drops. For SamPrivSyn and CALM, when the dimension increases, the number of each two-way marginal sampled

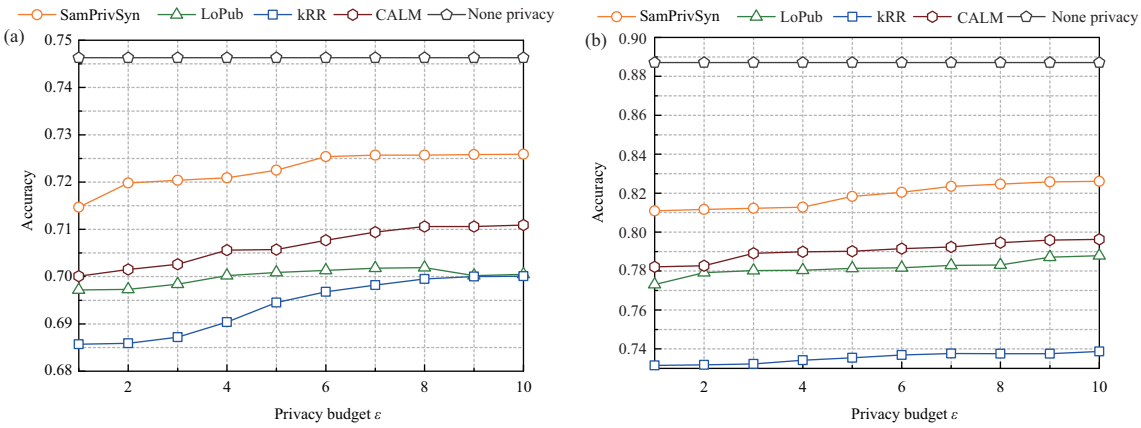

Figure 6 (Color online) Accuracy of logistic regression on the synthetic datasets generated by different LDP-enabled data synthesis methods. (a) Adult dataset; (b) Bank Marketing dataset.

to the data collector decreases. As a result, the estimation also drops. However, SamPrivSyn still performs much better than the other methods, regardless of the dimension size. To make the experimental results more convincing, this study also experimented with the real-world dataset Bank Marketing, with the results shown in Figure 5(e). It can be observed that the proposed SamPrivSyn method still outperforms other approaches. In summary, the above experiments prove the effectiveness of the proposed SamPrivSyn method in joint distribution estimation under LDP.

Logistic regression. To observe the overall performance of our proposed SamPrivSyn method, the performance of logistic regression was also evaluated on the synthetic datasets Adult and Bank Marketing generated by the four methods. The experimental results are shown in Figure 6, where Figure 6(a) shows the adult dataset, and Figure 6(b) shows the Bank Marketing dataset. The accuracy on the original datasets was evaluated, and the synthetic datasets were used to train the logistic model. As shown in Figures 6(a) and (b), the accuracy values (the gray line) for the original datasets Adult and Bank Marketing are 0.746 and 0.887, respectively. It can also be seen that the accuracy increases with an increase of the privacy budget, and the proposed SamPrivSyn method has great superiority compared with the other methods. This indicates that the proposed method is capable of better preserving the correlation between attributes, and the synthetic datasets are more similar to the original ones. Notably, the synthetic datasets are all generated with the same parameters in the estimation accuracy experiments. kRR still had the worst performance on logistic regression, with LoPub as second. Comparing Figure 6(a) with Figure 6(b), it can be observed that the accuracy on the synthetic Adult dataset is closer to the real accuracy because the dimension of Adult is smaller than that of Bank Marketing, and there is not much difference in the number of data records; therefore, the performance on the dataset Adult is better. In conclusion, based on the above results, it has been demonstrated that our proposed SamPrivSyn method is more effective in high-dimensional synthetic data under LDP.

5.2 Communication cost

Figure 7 shows the communication cost of all the methods, where the communication cost is defined as the whole bits that are sent to the server-side from the client-side. In this setting, the number of the data records was considered to be 30000, with the blue bar, gray bar, and orange bar representing LoPub, kRR, and SamPrivSyn methods, respectively. Specifically, Figure 7(a) shows that the communication cost varies with the dimension when the number of data records is set to a constant. It can be observed that the communication cost of LoPub is particularly high compared with the other methods, and SamPrivSyn is the smallest. Also, in Figure 7(b), the communication cost changes along with different data records on the four-dimensional dataset. Similarly, the communication cost of LoPub is still the highest, and that of SamPrivSyn is the smallest. Therefore, it can be concluded that both LoPub and kRR are not suitable for the high-dimensional setting, while SamPrivSyn can perform well, with low communication cost and high estimated accuracy.

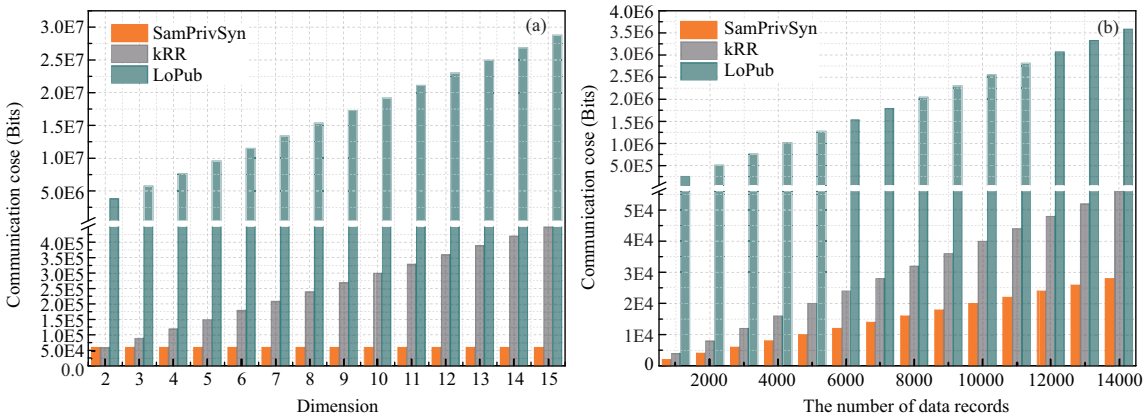

Figure 7 (Color online) Communication cost of different methods. The horizontal axes are (a) the dimension and (b) the number of the data records respectively, and the vertical axis is the communication cost.

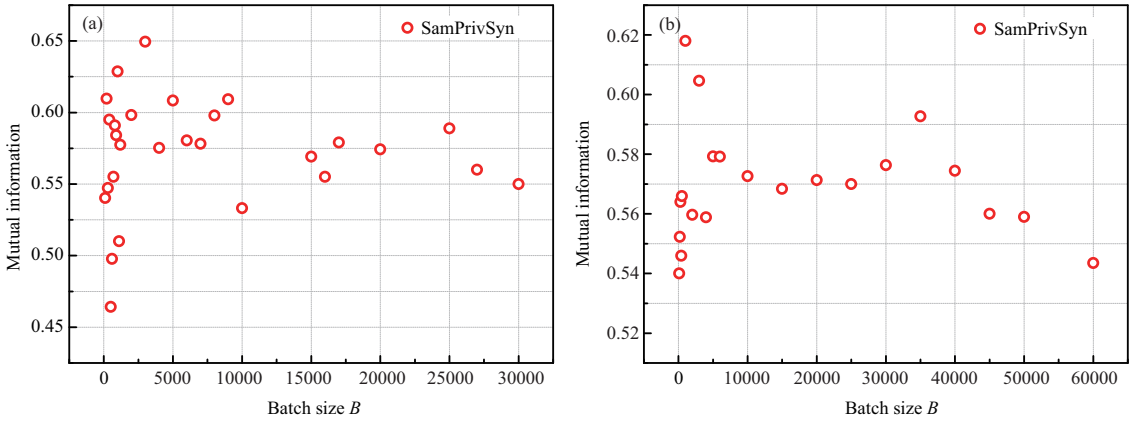

Figure 8 (Color online) Mutual information of {marital-status, relationship} with different batch sizes. (a) Adult; (b) SYN3-Adult.

5.3 Parameter sensitivity

Experiments were also conducted to observe the impact of different batch sizes B on performance. The privacy budget $\epsilon = 5$, and the batch size was changed to investigate the difference in correlation between the estimated distribution. The most correlated attribution pairs, marital-status and relationship, were selected to determine the impact of batch size on correlation. Figure 8 shows the experimental results on the Adult and SYN3-Adult datasets. The real mutual information in Adult and SYN3-Adult is approximately 0.71. At the beginning, with the increase in batch size, the mutual information of the estimated attributes becomes closer to the real value, but when it reaches the optimal value, the mutual information deviates from the real value in the overall trend with the increase in the batch size. This is because when the batch size is very small, the estimated distribution is inaccurate at the beginning such that the sampled records do not obey the distribution of mutual information. However, when the batch size is too big, the first sampled records according to the initialized mutual information have a greater impact on the sampled results. From the experimental results, the mutual information is closest to the real value when the budget batch size is between 1000 and 5000; however, it is difficult to give an accurate optimal batch size.

6 Conclusion

In this paper, a high-dimensional data synthesis method under LDP, called SamPrivSyn, was proposed. This method is composed of two modules: marginal sampling and data generation. In the marginal sampling phase, instead of sending a complete perturbed record to the server, which entails a huge communication cost and introduces large-scale noise when the dimension is very high, SamPrivSyn only

samples two attributes of one record based on the mutual information between attributes to preserve the correlations as much as possible. Each user then sends the sampled records to the server to construct two-way marginals. In the data generation phase, the server-side reconstructs the original dataset according to these marginals. Finally, extensive comparison experiments on real-world datasets demonstrated the superiority of SamPrivSyn.

Acknowledgements This work was supported by Strategic Research and Consulting Project of the Chinese Academy of Engineering (Grant No. 2022-XY-107).

References

1 Wang W, Xi J, Chen H. Modeling and recognizing driver behavior based on driving data: a survey. Math Problems Eng, 2014, 2014: 1–20
2 Preis T, Moat H S, Stanley H E. Quantifying trading behavior in financial markets using google trends. Sci Rep, 2013, 3: 1684
3 Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Conference on Security Symposium, 2014. 17–32
4 Ohlhorst F J. Big Data Analytics: Turning Big Data Into Big Money. Hoboken: John Wiley & Sons, 2012
5 Dwork C. Differential Privacy: A Survey of Results. Berlin: Springer, 2008
6 Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates. In: Proceedings of IEEE 54th Annual Symposium on Foundations of Computer Science, 2013
7 Nguyên T T, Xiao X K, Yang Y, et al. Collecting and analyzing data from smart device users with local differential privacy. 2016. ArXiv:1606.05053
8 Wang T, Li N, Jha S. Locally differentially private heavy hitter identification. IEEE Trans Dependable Secure Comput, 2019, 18: 982–993
9 Erlingsson Ú, Pihur V, Korolova A. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014. 1054–1067
10 Differential Privacy Team, Apple. Learning with privacy at scale. 2017. https://machinelearning.apple.com/research/learning-with-privacy-at-scale
11 Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, 2016. 2436–2444
12 Bassily R, Smith A. Local, private, efficient protocols for succinct histograms. In: Proceedings of the 47th ACM Symposium on Theory of Computing, 2015. 127–135
13 Ye M, Barg A. Optimal schemes for discrete distribution estimation under locally differential privacy. IEEE Trans Inform Theor, 2018, 64: 5662–5676
14 Xue Q, Zhu Y, Wang J. Joint distribution estimation and Naïve Bayes classification under local differential privacy. IEEE Trans Emerg Top Comput, 2021, 9: 2053–2063
15 Duchi J C, Jordan M I, Wainwright M J. Local privacy, data processing inequalities, and statistical minimax rates. 2013. ArXiv:1302.3203
16 Qin Z, Yang Y, Yu T, et al. Heavy hitter estimation over set-valued data with local differential privacy. In: Proceedings of ACM Sigsac Conference on Computer and Communications Security, 2016. 192–203
17 Ren X, Yu C M, Yu W, et al. LoPub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Trans Inform Forensic Secur, 2018, 13: 2151–2166
18 Warner S L. Randomized response: a survey technique for eliminating evasive answer bias. J Am Statistical Assoc, 1965, 60: 63–69
19 Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT Theor Comput Sci, 2014, 9: 211–407
20 Li N, Lyu M, Su D, et al. Differential privacy: from theory to practice. Synthesis Lectures Inf Security Privacy Trust, 2016, 8: 1–138
21 Mcsherry F, Talwar K. Mechanism design via differential privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), 2007
22 Wei J, Lin Y, Yao X, et al. Differential privacy-based genetic matching in personalized medicine. IEEE Trans Emerg Top Comput, 2021, 9: 1109–1125
23 Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we learn privately? SIAM J Comput, 2008, 40: 793–826
24 Kairouz P, Oh S, Viswanath P. Extremal Mechanisms for Local Differential Privacy. Cambridge: MIT Press, 2014
25 Wang T, Blocki J, Jha S K. Locally differentially private protocols for frequency estimation. In: Proceedings of the 26th USENIX Security Symposium, 2017
26 Zhang Z, Wang T, Li N, et al. CALM: consistent adaptive local marginal for marginal release under local differential privacy. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018. 212–229
27 Wang N, Xiao X, Yang Y, et al. Collecting and analyzing multidimensional data with local differential privacy. In: Proceedings of IEEE 35th Annual International Conference on Data Engineering (ICDE), 2019
28 Ye Q, Hu H, Meng X, et al. PrivKV: key-value data collection with local differential privacy. In: Proceedings of IEEE Symposium on Security and Privacy (SP), 2019. 317–331
29 Gu X, Li M, Cheng Y, et al. PCKV: locally differentially private correlated key-value data collection with optimized utility. In: Proceedings of the 29th USENIX Security Symposium, 2020. 967–984
30 Sun L, Zhao J, Ye X, et al. Conditional analysis for key-value data with local differential privacy. 2019. ArXiv:1907.05014
31 Cormode G, Kulkarni T, Srivastava D. Answering range queries under local differential privacy. Proc VLDB Endow, 2019, 12: 1126–1138
32 Wang T, Ding B, Zhou J, et al. Answering multi-dimensional analytical queries under local differential privacy. In: Proceedings of the International Conference on Management of Data, 2019. 159–176
33 Du L, Zhang Z, Bai S, et al. AHEAD: adaptive hierarchical decomposition for range query under local differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2021. 1266–1288
34 Zhang Z, Wang T, Honorio J, et al. PrivSyn: differentially private data synthesis. In: Proceedings of the 30th USENIX Security Symposium, 2021

Profile of Changjun JIANG

Prof. Changjun JIANG received his B.S. degree in computational mathematics and M.S. degree in computer software and theory from Shandong University of Science and Technology in 1986 and 1991, respectively. He obtained his Ph.D. degree in control theory and engineering from the Institute of Automation, Chinese Academy of Sciences in 1995. He did his postdoctoral research at the Institute of Computing Technology, Chinese Academy of Sciences from 1995 to 1997. From 1997 to 1998, he was a visiting researcher at the City University of Hong Kong. From 1986 to 1999, he worked at Shandong University of Science and Technology. Since 1999, he has been working at Tongji University as a professor and doctoral supervisor. During the period from 2015 to 2019, he was the president of Donghua University.

Currently, he is a chair professor at Tongji University, director of the Key Laboratory of Embedded Systems and Service Computing of the Ministry of Education, and director of the Cyber Finance Security Collaborative Innovation Center (co-constructed by the Ministry of Education and Shanghai). He also serves as vice chairman of Shanghai Association for Science and Technology, honorary professor of Brunel University London, vice chairman of China Cloud System Industry Innovation Strategic Alliance, chairman of China Artificial Intelligence Society, executive director of China Society of Automation, and a member of Shanghai Science and Technology Innovation Board Advisory Committee. He is a fellow of the British Institute of Engineering and Technology, the Chinese Society for Artificial Intelligence, and the Chinese Society of Automation. He is the deputy editor-in-chief of *Computing and Informatics*, and the editorial board member of *Big Data Mining and Analytics*, *Security and Safety*, *Journal of Computer*, *Journal of Software*, *Journal of Electronics*, etc. In 2021, he was elected a member of the Chinese Academy of Engineering.

Prof. Jiang has devoted himself to the research of network financial security for more than 30 years, and is a leader in this field in the country. He carried out in-depth and systematic research from three levels of basic theory, core technology and business system, formed a set of core methods and common technologies for online transaction risk control, and overcame the major technical problems of fast and accurate transaction risk prevention and control. He established the country's first online transaction risk prevention and control system and standards, making pioneering contributions to the country becoming an international "leader" in this field. His main contributions are as follows.

Creation of a behavioral theory of concurrent systems

The PN machine model of the concurrent language recognizer and the vector grammar of the concurrent language generator are systematically proposed, the PN machine theory of the network concurrent system is established, the behavioral energy level spectrum of the concurrent system is given, and the interaction and cooperation of the concurrent system are revealed. Based on the behavior correlation mechanism, the moment function method of quantitative analysis is proposed, and the formal method of flow decomposition is established, which solves the technical problem of concurrent decoupling of the system.

Invention of concurrent scheduling technology for network resource management and optimization

A dynamic double-matching algorithm suitable for independent task scheduling in heterogeneous environments is proposed, and the task execution time and completion time are comprehensively investigated to meet the goal of high timeliness. Cabin computing, a new computing mode, is proposed which provides cross-domain resource configuration and collaborative computing integration environments for the entire life cycle of IT tasks, and solves the technical problems of configuration uncertainty and timeliness of concurrent computing environments.

He is the first in the world to propose a behavioral certification method for risk prevention and control

A behavior authentication mechanism based on the analytical model is designed, and a legal transmission sequence determination algorithm based on the synchronous behavior set is proposed, which breaks through the bottleneck of the poor real-time performance of the existing legal transmission sequence enumeration and determination. A hierarchical diagnosis and treatment and hierarchical risk control mechanism is established. The indicator-related logic analysis and data quantification model calculation are integrated and embedded. A hierarchical intelligent diagnosis and treatment method of transaction risk general inspection-specialized diagnosis-collective diagnosis is proposed.

His research studies are serving the country's anti-financial fraud, effectively supporting the Yunjian 2020 campaign of the Ministry of Public Security, and have been highly praised by national regulatory agencies. The research studies serve the transaction security of more than 900 million Alipay real-name users in more than 200 countries and regions around the world. The transaction compensation rate is 5% per million, which is far lower than that at the US PayPal, the best international online payment platform. His research work has also served more than 500 units, such as the Industrial and Commercial Bank of China and the Shanghai Free Trade Zone.

The research results have been positively evaluated and cited many times by well-known experts such as academicians from the United States, the United Kingdom, Germany, Sweden, India, and other countries. He has obtained more than 140 invention patents from China, the United States, Germany, and other countries, and 22 international

PCTs. He presided over and participated in the formulation of 18 national and industry standards; published more than 300 papers (including 82 papers in ACM/IEEE Transactions) and 5 monographs in Chinese and English. He has won the Second Prize in the National Technological Invention Award (ranked 1), the Second Prize twice in the National Science and Technology Progress Award (both ranked 1), and 8 first prizes at provincial and ministerial levels.

In addition, he has successively won the HO PAN CHING YI Award (1996) in the field of discrete event dynamic systems (DEDS), the first national 100 outstanding doctoral dissertations (1999), the Shanghai Dawning Program (2000) and its tracking Program (2008), University Young Teacher Award (2001), National Science Fund for Distinguished Young Scholars (2001), Teaching and Research Award Program for Outstanding Young Teachers of the Ministry of Education (2001), Shanghai Outstanding Discipline Leader Program (2004), Head of the Excellent Innovation Team of the Ministry of Education (2007), Shanghai Leading Talent Program (2009), Chief Scientist of the 973 Project (2009), National Outstanding Scientific and Technological Worker (2016), National Innovation Award (2020).

Selected publications

• Jiang C J. Vector grammars and PN machines. Sci China Ser E-Tech Sci, 1996, 39: 50–60

• Jiang C J. PN Machine Theory of Discrete Event Dynamic Systems (in Chinese). Beijing: Science Press, 2000

• Jiang C J. Behavior theory and applications of Petri net (in Chinese). Beijing: Higher Education Press, 2003

• Jiang C J. Polynomial-time algorithm for the legal firing sequences problem of a type of synchronous composition Petri nets. Sci China Ser F-Inf Sci, 2001, 44: 226–233

• Wang H Q, Jiang C J, Liao S Y. Behaviour relations in synthesis process of Petri net models. IEEE Trans Robot Autom, 2000, 16: 400–407

• Zhi Q, Jiang C J. A scheduling algorithm suitable for heterogeneous computing environment (in Chinese). Acta Autom Sin, 2005, 31: 865–872

• Wang S, Ding Z J, Jiang C J. Elastic scheduling for microservice applications in clouds. IEEE Trans Parallel Distrib Syst, 2021, 32: 98–115

• Jiang C J, Ding Z J, Yu J, et al. Cabin computing (in Chinese). Sci Sin Inform, 2021, 51: 1233–1254

• Jiang C J, Wang P W. Networking Computing (in Chinese). Beijing: Science Press, 2020

• Jiang C J, Ding Z J, Wang J L, et al. Big data resource service platform for the internet financial industry. Chin Sci Bull, 2014, 59: 5051–5058

• Jiang C J, Yu W Y. Risk Control Theory of Online Transaction (in Chinese). Beijing: Science Press, 2018

• Jiang C J, Song J H, Liu G J, et al. Credit card fraud detection: a novel approach using aggregation strategy and feedback mechanism. IEEE Internet Things J, 2018, 5: 3637–3647

• Jiang C J, Wang J L. Intelligence originating from human beings and expanding in industry — a view on the development of artificial intelligence. Strategic Study Chin Academy Eng, 2018, 20: 93–100

• Jiang C J, Fang Y, Zhao P H, et al. Intelligent UAV identity authentication and safety supervision based on behavior modeling and prediction. IEEE Trans Indust Inform, 2020, 16: 6652–6662