Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 7846547, 16 pages
https://doi.org/10.1155/2018/7846547

Research Article

Using MOPSO for Optimizing Randomized Response Schemes in Privacy Computing

Zhiqiang Gao, Xiaolong Cui, Yanyu Duan, Zhang Jun, and Zhensheng Peng

Department of Information Engineering, Engineering University of Chinese People's Armed Police Force, Xi'an, China

Correspondence should be addressed to Zhiqiang Gao; 1090398464@qq.com

Received 17 November 2017; Revised 6 February 2018; Accepted 22 February 2018; Published 3 April 2018

Academic Editor: Khaled Loukhaoukha

It is a challenging concern in data collecting, publishing, and mining when personal information is controlled by untrustworthy cloud services with unpredictable risks for privacy leakages. In this paper, we formulate an information-theoretic model for privacy protection and present a concrete solution to theoretical architecture in privacy computing from the perspectives of quantification and optimization. Thereinto, metrics of privacy and utility for randomized response (RR) which satisfy differential privacy are derived as average mutual information and average distortion rate under the information-theoretic model. Finally, a discrete multiobjective particle swarm optimization (MOPSO) is proposed to search optimal RR distorted matrices. To the best of our knowledge, our proposed approach is the first solution to optimize RR distorted matrices using discrete MOPSO. In detail, particles' position and velocity are redefined in the problem-guided initialization and velocity updating mechanism. Two mutation strategies are introduced to escape from local optimum. The experimental results illustrate that our approach outperforms existing state-of-the-art works and can contribute optimal Pareto solutions of extensive RR schemes to future study.

1. Introduction

With the rapid development of artificial intelligence [1, 2] and mobile computing, big data has conceived booming series of services and has been regarded as a ubiquitously fundamental resource. New paradigms such as smart city [3], Takeout [4], and mobile pay [5] are growing at an amazing speed. Undoubtedly, these applications can provide users with accurate and personalized services with great convenience. However, large amounts of users' personal information, including consuming habits, income status, and location, are collected beyond the control of their real owners. Especially, massive privacy concerns, including AOL search logs scandal (2006) [6, 7], de-anonymity and cancellation of Netflix prize (2009) [8, 9], complaint of privacy settings by Facebook (2009) [10], privacy breach of New York taxi trips (2016) [11], and massive breach of Equifax data (2017) [12] have tightened the whole society's nerve over sensitive data collecting, releasing, and mining. It can be regarded as a profound privacy challenge and a motivating research field in privacy computing. On the whole, categories of privacy computing can be divided into three imperative aspects: privacy preserving data collecting (PPDC), privacy preserving data mining (PPDM), and privacy preserving data publishing (PPDP).

Fruitful works of anonymity-based traditional privacy preserving methods [13–17], as well as randomization-based approaches [18–20], have been prominently studied and extensively applied in so many fields ranging from social networks [21], location-based services [22], and smart healthcare [23–25] to intelligent transportation [26]. However, most of traditional approaches are designed on specific application scenarios. Notably, newly arising challenges in privacy computing can be summarized into the following four urgent issues: (1) there still lack rigorous theoretical architecture and fundamental principles of privacy computing, which can systematically quantify privacy and describe relationship between protection level and utility loss; (2) large amounts of personal data are collected by untrustworthy collector, which enlarges the gap between owners and collectors in risk of privacy leakages; (3) new computing paradigms, such as MapReduce [27], Storm [28], and Spark [29], and deep learning, are in need of further security enhancement from the perspective of privacy computing; (4) optimization of the trade-off between individual privacy and data utility in

privacy computing schemes should be further studied under the premise of satisfying users' diversified requirements.

In this paper, our work is concentrated on the most challenging issue of quantifying and optimizing the trade-off between data privacy and utility in privacy computing. Derived from the concept of privacy computing, randomized response [30, 31] is highly efficient in PPDC and capable of masking private data, while maintaining the reconstruction ability of aggregate information with tolerable errors. In RR schemes, individuals' private data is transformed by RR distorted matrices (for ease of illustration, we use RR matrices for short.) to nonsensitive disguised data. Although RR schemes have been widely studied in sensitive survey, the primary problem of searching for optimal RR matrices is rarely explored. Additionally, the search space of RR matrix is astronomical and infeasible to exploit with brute-force method. We transform the comparison of RR schemes into the model of multiobjective optimization problem, for which multiobjective particle swarm optimization (MOPSO) [32] is extremely fit to provide a much more diversified set of optimal Pareto fronts than other solutions. Thus, different RR schemes can be compared in a quantified manner by two conflicting metrics of privacy and utility under the information-theoretic model. Moreover, differential privacy can provide a rigorous and fundamental guarantee for the optimization of RR schemes in privacy computing. Overall, MOPSO is utilized to search optimal RR matrices by the metrics of privacy and utility derived from the information-theoretic model. Our experimental results show a satisfying improvement over existing works and a wide range of optimal Pareto fronts are obtained for extensive schemes. The contributions of our work can be summarized as follows:

(i) We formulate the information-theoretic model for privacy quantification and present a solution to theoretical architecture in privacy computing.

(ii) We derive metrics of privacy and utility under the information-theoretic model for RR matrices which satisfy differential privacy.

(iii) We proposed a specified discrete MOPSO to find the optimal set of RR matrices under two conflicting goals: privacy and utility.

The remainder of our paper is structured as follows. Section 2 makes a systematic review on related works in privacy computing from the perspectives of quantification and optimization in PPDC, PPDP, and PPDM. In Section 3, the information entropy model in privacy computing is illustrated and the multiobjective optimization problem for quantifying privacy and data utility of RR schemes under ε-differential privacy is modeled. Section 4 presents a solution to searching optimal set of RR matrices by discrete MOPSO and Section 5 discusses the experimental results. Conclusion and future work are provided in Section 6.

2. Related Work

Roughly speaking, there can be three typical scenarios in privacy computing, namely, data collecting, data publishing, and data mining according to the convoluted life-cycle of private data. In fact, privacy computing (will be detailed in Section 3) emphasizes the quantification and optimization.

Privacy preserving data collection (PPDC), regarded as local privacy, is a strategy that perturbs the data locally before they reach the untrustworthy data-collector. The goal of PPDC is to satisfy accurate estimation of population statistics as well as guarantee the privacy of individual simultaneously. Wang et al. [33] compared randomized response with Laplace mechanism using differential privacy in PPDC and recommended a RR-based scheme with less utility loss. RAPPOR in [31] is also a RR-based application of differential privacy in Chrome, which allows crowd-sourcing statistics from client-end browsers with rigorous ε-differential privacy guarantee. PPDC can achieve meaningful aggregate information inferences while preserving the privacy of client-side users. Works in [30, 34–37] can be categorized into the scenario where data-collector wishes to grasp the distribution of original data while each client is just required to submit a perturbed version. Intuitively, these methods are in similarity with reconstruction on noised data. Notably, randomized response and local differential privacy [38] can be two powerful tools in PPDC. Coincidentally, randomized response can satisfy ε-differential privacy by specified parameter settings. However, FRAPP in [39] just generalized each element in distorted matrix by a single metric of accuracy. While, Agrawal and Srikant [36] just measured data reconstruction under differential privacy in terms of data utility by mean square error in the process of data mining. Huang and Du proposed a scheme of OptRR [34] which combines SPEA2 to find optimal distorted matrices in a heuristic way. Unfortunately, they mistake the dominance relationship of Pareto solutions. Additionally, the upper bound of privacy that derived from maximum a posteriori (MAP) estimate is not rigorous compared with differential privacy budget. In our work, two metrics of individuals' privacy and data utility are optimized simultaneously under a unified model of multiobjective optimization problem, which can present a much more thorough illustration of RR matrices than FRAPP. Additionally, discrete MOPSO is proposed to provide more diversified optimal solutions than SPEA2. Particles' position and velocity are both redefined and a novel velocity updating mechanism is adopted with two mutation strategies. The experiments show that our approach outperforms SPEA2 and FRAPP.

Intuitively, there is no prominent discrimination between privacy preserving data publishing (PPDP) and privacy preserving data mining (PPDM) for the reason that the two models are built on the same assumption of trusted data-curator. Also, privacy preserving data collection can be covered by PPDM when collected data is directly served for data mining. According to the life-cycle of data and typical appliances, PPDC can be studied independently. On the one hand, there are two settings, namely, interactive and noninteractive in PPDP. Thereinto, k-anonymity [13] and its variants [14–17] are extensively studied on the assumption that data can be divided into sensitive attributes and quasi-identifiers in tabular forms. However, linkage attack, differential attack, and background knowledge attack are proposed in

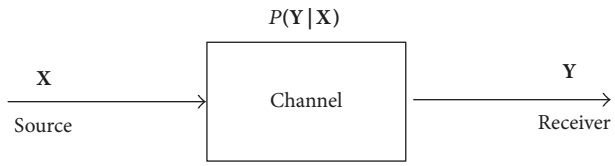


Figure 1: General channel model of information theory.

sequence to destroy the above privacy preserving models. To this end, anonymity-based approaches fail to provide strong privacy guarantees compared with the rigorous definition of differential privacy. Li et al. [40] preceded with randomized sampling before k-anonymity to achieve (ε, δ)-differential privacy, which shed lights on the broken gap between semantic security and syntactic security in PPDP. On the other hand, privacy preserving data mining was first proposed by Agrawal and Srikant in 2000 [36]. Later, Agrawal and Aggarwal [35] utilized EM algorithm to preserve individual privacy in the reconstruction of data distribution. In [30], Du and Zhan proposed a RR-based scheme to build decision trees on perturbed data and Agrawal and Haritsa presented a solution of privacy preserving computation for perturbed data with multiple attributes from multiple clients in [39].

Although several lines of privacy computing fall into the scope of PPDC, PPDP, and PPDM, there still lacks a systematical architecture of privacy computing with fundamental supports in rigorous and provable theory. Our work systematically illustrates the domain of privacy computing and builds a new paradigm of multiobjective optimization problem, which is aimed at quantifying individuals' privacy and data utility in PPDC. Guaranteed by differential privacy in randomized response, discrete MOPSO presents an excellent solution to the trade-off between privacy and data utility in optimal RR matrices and performs better than the state-of-the-art schemes.

3. Privacy Computing

3.1. Information-Theoretic Model. Privacy computing [41] can be defined as an open theoretic system, which is comprised of computing models, application scenarios, quantitative definitions, and powerful privacy preserving techniques. We quantify the concept of privacy under the general definition of information theory rather than terms that are restricted to sensitive or personal data from laws and regulations.

Inspired by the process of information transferring in communication system, intrinsic similarity can be applied in privacy computing. Information will be affected by a noisy channel which can be regarded as a beneficial protection when they are transferred to the receiver-side. Here, we review information theory basics before demonstrating how to quantify privacy and utility in privacy computing.

Information theory [42] is a mathematical framework for quantifying information transmission in communication systems. Figure 1 illustrates the scenario of a general channel model between source and receiver, which is illustrated in Scenario 1.

Scenario 1. The probability space is comprised of discrete memoryless information source $\mathbf{X}$, receiver $\mathbf{Y}$, and channel $P(\mathbf{Y} \mid \mathbf{X})$ as follows:

$$\begin{bmatrix} \mathbf{X} \\ P(\mathbf{X}) \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_s \\ P(a_1) & P(a_2) & \cdots & P(a_s) \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{Y} \\ P(\mathbf{Y}) \end{bmatrix} = \begin{bmatrix} b_1 & b_2 & \cdots & b_r \\ P(b_1) & P(b_2) & \cdots & P(b_r) \end{bmatrix},$$

$$P(\mathbf{Y} \mid \mathbf{X}) = \begin{bmatrix} p(b_1 \mid a_1) & p(b_2 \mid a_1) & \cdots & p(b_r \mid a_1) \\ p(b_1 \mid a_2) & p(b_2 \mid a_2) & \cdots & p(b_r \mid a_2) \\ \vdots & \vdots & & \vdots \\ p(b_1 \mid a_s) & p(b_2 \mid a_s) & \cdots & p(b_r \mid a_s) \end{bmatrix}, \quad (1)$$

$$\text{subject to} \quad \sum_{i=1}^{s} P(a_i) = 1, \quad 0 \le P(a_i) \le 1$$

$$\sum_{i=1}^{r} P(b_i) = 1, \quad 0 \le P(b_i) \le 1$$

$$\sum_{j=1}^{r} P(b_j \mid a_i) = 1,$$

$$0 \le P(b_j \mid a_i) \le 1, \ i \in \{1, 2, \ldots, s\}.$$

Source $\mathbf{X} \in \{a_1, a_2, \ldots, a_s\}$ and receiver $\mathbf{Y} \in \{b_1, b_2, \ldots, b_s\}$ (sometimes $r \neq s$) are discretized into mutually exclusive and exhaustive categories with corresponding probability as shown in Abbreviations. From the perspective of receiver's observations, information source is an abstract mathematical representation for a physical entity that produces a succession of discrete symbols in a randomized manner. $P(\mathbf{Y} \mid \mathbf{X})$ is the feature of channel with distortion noise. Overall, the transmission flow can be formulated as

$$P^T(\mathbf{Y}) = P(\mathbf{Y} \mid \mathbf{X}) P^T(\mathbf{X}). \quad (2)$$

Information theory quantifies how much information a receiver $\mathbf{Y}$ carries about the source $\mathbf{X}$, which is applicable in the scenario of measuring how much original information can be estimated from distorted data. Applying Shannon's mathematical theory in information theory to privacy computing, we can derive several quantitative representations of information.

Pertinent to information source in probability space, self-information $\mathbf{I}(a_i)$ denotes the uncertainty of event a_i before the following occurs:

$$\mathbf{I}(a_i) = \log \frac{1}{P(a_i)}. \quad (3)$$

Prior entropy of $\mathbf{H}(X)$ denotes the average uncertainty about $\mathbf{X}$ before received and the probability of secret leakage. The Shannon entropy of $\mathbf{X}$ is defined as

$$\mathbf{H}(X) = \sum_{i=1}^{r} P(a_i) \log \frac{1}{P(a_i)}. \quad (4)$$

Posterior entropy $\mathbf{H}(\mathbf{X} \mid \mathbf{Y})$ denotes the average uncertainty about $\mathbf{X}$ after being received and the chance of attackers detecting secrets over the output

$$\begin{aligned}\mathbf{H}(\mathbf{X} \mid \mathbf{Y}) &= \mathbf{E}\left(\mathbf{H}\left(\mathbf{X} \mid b_j\right)\right) = \sum_{j=1}^{s} P\left(b_j\right) \mathbf{H}\left(\mathbf{X} \mid b_j\right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{s} P\left(a_i b_j\right) \log \frac{1}{P\left(a_i \mid b_j\right)} \\ &= \sum_{X,Y} P(xy) \log \frac{1}{P(x \mid y)}.\end{aligned} \tag{5}$$

Mutual information $\mathbf{I}(\mathbf{X}; \mathbf{Y})$ denotes how much $\mathbf{X}$ can be preserved from $\mathbf{Y}$ and gains by observing information leakage

$$\mathbf{I}(\mathbf{X}; \mathbf{Y}) = \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X} \mid \mathbf{Y}). \tag{6}$$

Average mutual information can be calculated as

$$\begin{aligned}\mathbf{I}(\mathbf{X}; \mathbf{Y}) &= \mathop{\mathbf{E}}_{\mathbf{X},\mathbf{Y}}\left[\mathbf{I}(x; y)\right] = \sum_{\mathbf{X},\mathbf{Y}} P(xy)\, I(x; y) \\ &= \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{P(x \mid y)}{P(x)}.\end{aligned} \tag{7}$$

On the other hand, the decoding function $\mathbf{F}$ can be used as the inverse process of channel when b_j received from source a_i

$$\mathbf{F}\left(b_j\right) = a_i. \tag{8}$$

When signal b_j is received, conditional error rate P_E can be measured as follows:

$$\begin{aligned}P\left(e \mid b_j\right) &= 1 - P\left(a_i \mid b_j\right) = 1 - P\left[\mathbf{F}\left(b_j\right) \mid b_j\right]. \\ P_E &= \sum_{\mathbf{Y}} P\left(b_j\right) P\left(e \mid b_j\right) \\ &= \sum_{\mathbf{Y}} P\left(b_j\right)\left\{1 - P\left[\mathbf{F}\left(b_j\right) \mid b_j\right]\right\} \\ &= 1 - \sum_{\mathbf{Y}} P\left[\mathbf{F}\left(b_j\right) b_j\right] \\ &= \sum_{\mathbf{X},\mathbf{Y}} P\left(a_i b_j\right) - \sum_{\mathbf{Y}} P\left[\mathbf{F}\left(b_j\right) b_j\right] \\ &= \sum_{\mathbf{X},\mathbf{Y}} P\left(a_i b_j\right) - \sum_{\mathbf{Y}} P\left(a^* b_j\right) \\ &= \sum_{\mathbf{Y},\mathbf{X}-a^*} P\left(a_i b_j\right).\end{aligned} \tag{9}$$

It is worth noting that privacy has many similarities with information theory and anonymity from a mathematical point of view. Thus, privacy can be generalized and formulated as follows.

Definition 2 (privacy). Given i.i.d vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and the probability $P(x_i)$, privacy of $\mathbf{X}$ can be defined by information entropy in probability space

$$\text{privacy}(\mathbf{X}) = -\sum_{n} p(x_i) \log p(x_i). \tag{10}$$

When variables of $\mathbf{X}_i$ and $\mathbf{X}_j$ are not independent, mutual information can be adopted to measure the information contained in one process about another through a noisy channel environment

$$\text{privacy}\left(\mathbf{X}_i, \mathbf{X}_j\right) = \sum p\left(x_i, x_j\right) \log \frac{p\left(x_i \mid x_j\right)}{p\left(x_i\right)}. \tag{11}$$

In practice, privacy can be estimated from distorted data by certain priori knowledge. Data utility can be measured by mean square error (MSE) of estimator $P(\widehat{\mathbf{X}})$ and $P(\mathbf{X})$ [34]

$$\text{MSE}(\mathbf{X}) = \mathbf{E}\left(P\left(\widehat{\mathbf{X}}\right) - P(\mathbf{X})\right)^2. \tag{12}$$

Equipped with the information-theoretic model of privacy above, we establish an open architecture of privacy computing with 7 key components, named tuple ($\mathbf{X}$, $\mathbf{P}$, $\mathbf{M}$, Φ, $\mathbf{\Omega}$, $\mathbf{T}$, $\mathbf{\Delta}$) as shown in Abbreviations.

In detail, the distribution of $\mathbf{X}$ plays a significant role in reconstruction and estimate of original attributes for private data. $\mathbf{P}$ denotes the set of participants, including owners of private information, analysts, attackers with arbitrary background knowledge, and curators and collectors of private information. $\mathbf{M}$ represents the metrics of quantifying privacy, for example, privacy and utility. Φ involves constraints on time and scenarios, such as data collection, data releasing, and data mining. $\mathbf{\Omega}$ can be extended to the operations covering the whole life-cycle of privacy, including PPDC, PPDP, and PPDM. $\mathbf{T}$ represents mature privacy computing paradigms, such as randomized response, differential privacy, and k-anonymity. Additionally, $\mathbf{\Delta}$ is the supplementary set of privacy computing. Theoretically, we can formulate the relationship of privacy and utility in the system of privacy computing

$$\begin{aligned}\text{privacy} &= \mathbf{f}(\mathbf{X}, \mathbf{P}, \mathbf{M}, \Phi, \mathbf{\Omega}), \\ \text{utility} &= \mathbf{g}(\mathbf{X}, \mathbf{P}, \mathbf{M}, \Phi, \mathbf{\Omega}).\end{aligned} \tag{13}$$

In summary, privacy and utility are quantified by metric functions of $\mathbf{f}$ and $\mathbf{g}$ with tuple of input parameter $\mathbf{X}$, $\mathbf{P}$, $\mathbf{M}$, Φ, and $\mathbf{\Omega}$ (the detailed description of $\mathbf{f}$ and $\mathbf{g}$ will be presented in Section 3.3). Specifically, in the concrete setting of our work, $\mathbf{X}$ denotes the collected categorical data from distributed clients. $\mathbf{P}$ represents the client-side user and the server-side untrustworthy collector. $\mathbf{M}$ is the evaluation metric of privacy and utility. Φ depicts the scenario setting where data collection occurs. $\mathbf{\Omega}$ is the operation of randomized response in PPDC. Notably, $\mathbf{T}$ denotes the paradigm of randomized response guaranteed by differential privacy (the concrete usage of Φ and $\mathbf{T}$ will be described in Section 4).

Intuitively, the scope of privacy computing involved in the life-cycle of privacy can be illustrated in Figure 2. When privacy is collected, published, mined, or stored without protection, it is no better than destroying them directly.

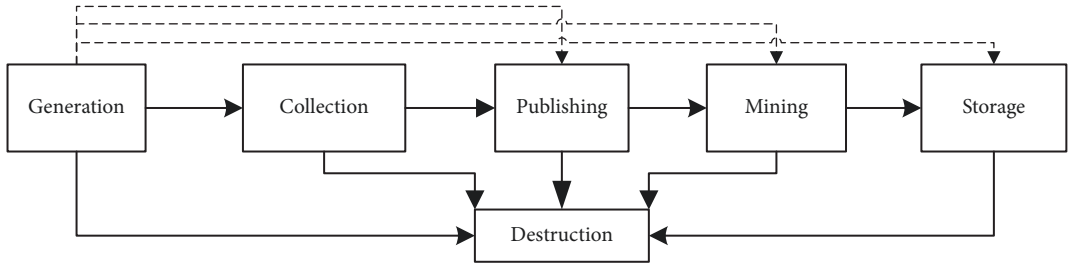


Figure 2: Life-cycle of privacy.

3.2. Randomized Response under Differential Privacy. Randomized response aims at eliciting information on sensitive survey and can be used to derive unbiased estimate from untruthful responses. Typically, aggregate information and relatively accurate reconstruction from distorted responses can be estimated and built while protecting local individual privacy by RR under differential privacy.

Treated as the most rigorous privacy preserving tool, differential privacy (DP) [43, 44] is widely studied in privacy computing. DP can provide effective protection over the outcome no matter how one opts in or opts out of the database.

Definition 3 (ε-differential privacy [18]). Given two statistical databases D and D' which satisfy $|D - D'| = 1$ (Hamming distance), randomized function $\mathbf{A}$ achieves ε-differential privacy on condition that

$$e^{-\varepsilon} \leq \frac{\Pr\left(\mathbf{A}\left(D\right) \in \mathbf{R}\right)}{\Pr\left(\mathbf{A}\left(D'\right) \in \mathbf{R}\right)} \leq e^{\varepsilon}, \tag{14}$$

where $\mathbf{A}(D)$ denotes the output within the domain of $\mathbf{R}$. ε is privacy budget which can balance privacy preserving level and utility of outputs.

It is well accepted that Laplace mechanism can calibrate i.i.d noise from Laplace distribution which satisfies differential privacy. However, compared with randomized response, MSE of Laplace mechanism is ill-performed under the single metric of data utility [33]. Thus, our work emphasizes randomized response which conducts sensitive categorical data collection in privacy computing as shown in Scenario 4 and Figure 3.

Scenario 4. Given that n individual $U_i \in \mathbf{U}$, $\mathbf{U} = \{U_1, U_2, \ldots, U_n\}$, with categorical private data $x_i \in \mathbf{X}$, $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, to protect data privacy, each individual U_i submits another distorted version $y_i \in \mathbf{X}$ within the same category domain to the untrustworthy data-collector by designed distorted matrices in randomized response.

Given that $\mathbf{P}(x_i)$ denotes the proportion of category x_i in the original sensitive data and $\mathbf{P}^*(x_i)$ denotes the proportion of x_i in the distorted data and given

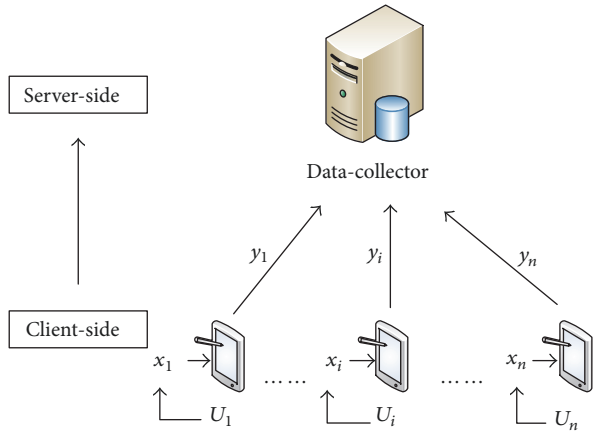


Figure 3: Randomized response scheme.

that $\overrightarrow{\mathbf{P}}_y^* = (P^*(x_1), P^*(x_2), \ldots, P^*(x_n))^T$ and $\overrightarrow{\mathbf{P}}_x = (P(x_1), P(x_2), \ldots, P(x_n))^T$, our designed distorted matrix $\mathbf{M}$ is as follows:

$$\mathbf{M} = \begin{pmatrix} p_{x_1x_1} & \cdots & p_{x_1x_n} \\ \vdots & \ddots & \vdots \\ p_{x_nx_1} & \cdots & p_{x_nx_n} \end{pmatrix}, \tag{15}$$

where $p_{yx} = p(y_i = y \mid x_i = x)$ denotes the probability that when original data input is x and disguised output is y, we can get the following equation:

$$\overrightarrow{\mathbf{P}}_y^* = \mathbf{M}\overrightarrow{\mathbf{P}}_x. \tag{16}$$

Compared with Scenario 1, we can bridge the gap between information theory and RR scheme in Scenario 4 of privacy computing. From the perspective of data-collector which is equal to the role of receiver in information-theoretic model, the number of categories x_i in distorted data N_i ($i = 1, \ldots, n$) and the total number of records N can be used to derive the maximum likelihood estimation (MLE) of $\widehat{\overrightarrow{\mathbf{P}}}_y^*$

$$\widehat{\overrightarrow{\mathbf{P}}}_y^* = \left(\frac{N_1}{N}, \ldots, \frac{N_n}{N}\right)^T. \tag{17}$$

In [37], Agrawal et al. proposed an iterative approach to estimate $\widehat{\overrightarrow{\mathbf{P}}}_x$, resulting however in high running time

consumption. Thus, we derive $\widehat{\overrightarrow{\mathbf{P}}}_x$ by matrix transformation based on (16). In other words, MLE of $\overrightarrow{\mathbf{P}}_x$ can be obtained when $\mathbf{M}$ is invertible. We regard it as a constraint of $\mathbf{M}$ in the case that determinant of $\mathbf{M}$ is nonzero. It can be concluded that $\mathbf{M}$ is the key factor that determines the accuracy of estimating $\overrightarrow{\mathbf{P}}_x$. Undoubtedly, as the amount of data grows, the distribution reconstruction of $\overrightarrow{\mathbf{P}}_x$ will be relatively close to the true distribution of original data

$$\widehat{\overrightarrow{\mathbf{P}}}_x = \mathbf{M}^{-1}\widehat{\overrightarrow{\mathbf{P}}}_y^{*}. \tag{18}$$

However, how to design distorted matrices and compare which matrix is optimal remain to be fundamentally unsolved. In [33], designed distorted matrix with larger diagonal elements is deemed to achieve better data utility while thoroughly ignoring the metric of privacy and failing to provide a diversified set of optimal distorted matrices. Then we take a concrete example of binary data collection for better understanding [45].

Example 5. Given $\mathbf{X} = \{x_1, x_2\}$ and 2×2 distorted matrix $\mathbf{M}$, the unbiased estimator and variance of the binary attribute x_1 are as follows:

$$\begin{aligned}
\mathbf{M} &= \begin{pmatrix} p_{x_1x_1} & p_{x_1x_2} \\ p_{x_2x_1} & p_{x_2x_2} \end{pmatrix}, \\
\widehat{\pi} &= \frac{p_{x_1x_1} - 1}{2p_{x_1x_1} - 1} + \frac{N_1}{\left(2p_{x_1x_1} - 1\right)N}, \\
\operatorname{Var}\widehat{\pi} &= \frac{1/4 - (\widehat{\pi} - 1/2)^2}{N} \\
&\quad + \frac{1/16\left(p_{x_1x_1} - 1/2\right)^2 - 1/4}{N}.
\end{aligned} \tag{19}$$

Derived from differential privacy in Definition 3, we can reach the conclusion that while $\max\{p_{x_1x_1}/p_{x_1x_2}, p_{x_2x_2}/p_{x_2x_1}\} \leq e^{\varepsilon}$ is satisfied, ε-differential privacy can be achieved. When $p_{x_1x_1} = e^{\varepsilon}/(e^{\varepsilon} + 1)$ and $p_{x_1x_2} = 1/(e^{\varepsilon} + 1)$, the designed distorted matrix can achieve the optimal utility [33]. Additionally, when a participant responses with a deniable probability of 1/4 in RR survey, we can get $\varepsilon = \ln(0.75/(1 - 0.75)) = \ln(3)$. That is, $\ln(3)$-differential privacy can guarantee the process of the sensitive information survey [31].

However, while obtaining the highest data utility, privacy is not perfectly protected. Meanwhile, although guaranteed by differential privacy, different designed distorted matrices, such as Warner's scheme in [45], Uniform Perturbation in [37], and FRAPP in [39], are difficult to determine which is optimal. Notably, the relationship of the two conflicting objectives can be defined as follows. Let $\mathbf{f}(\mathbf{M})$ and $\mathbf{g}(\mathbf{M})$ denote privacy and utility of RR matrices, respectively.

Definition 6 (Pareto dominance). $\mathbf{M}^*$ is in Pareto dominance i.f.f

$$\begin{aligned}
&\left(\forall \mathbf{F}_p \in \{\mathbf{f}(\mathbf{M}), \mathbf{g}(\mathbf{M})\} : \mathbf{F}_p(\mathbf{M}^*) \leq \mathbf{F}_p(\mathbf{M})\right) \\
&\quad \wedge \left(\exists \mathbf{F}_k \in \{\mathbf{f}(\mathbf{M}), \mathbf{g}(\mathbf{M})\} : \mathbf{F}_k(\mathbf{M}^*) < \mathbf{F}_k(\mathbf{M})\right).
\end{aligned} \tag{20}$$

Definition 7 (optimal RR matrix set). The set of all $\mathbf{M}^*$ in Pareto dominance from feasible domain $\mathbf{X}_p$ are

$$\text{Archive} = \left\{\mathbf{M}^* \mid \neg\exists \mathbf{M} \in \mathbf{X}_p : \mathbf{M} \succ \mathbf{M}^*\right\}. \tag{21}$$

Definition 8 (optimal RR matrix front). The set of objective functions obtained by Pareto optimal set are

$$\mathbf{P}_f = \left\{\left(\mathbf{f}(\mathbf{M}^*), \mathbf{g}(\mathbf{M}^*)\right) \mid \mathbf{M}^* \in \text{Archive}\right\}. \tag{22}$$

Obviously, optimal RR matrices are not unique and, actually, we can find a set of optimal RR matrices for further selection in search space. In the next subsection, we are focused on the quantification scheme for metrics of privacy and utility pertinent to $\mathbf{f}(\mathbf{M})$ and $\mathbf{g}(\mathbf{M})$.

3.3. Scheme for Quantification. Measuring privacy and utility in RR schemes is crucial for many reasons [46, 47]. (1) It is difficult to justify RR matrices just by investigator's intuition and experience. (2) The two criteria are conflicting in nature: when data utility increases, privacy will undoubtedly decrease to a certain degree and vice versa. However, combination of the two metrics can give thoroughly reasonable quantification schemes than a single one.

Meanwhile, information theory can quantify how much information a receiver $\mathbf{Y}$ carries about the source $\mathbf{X}$. Noisy channel can link source coding and receiver decoding by quantitative notions of average mutual information and distortion rate. In OptRR [34], privacy is measured by the amount of individual privacy estimated from the distorted data, and utility is measured by MSE of original data and the estimator. Differently and more convincingly, we derive metrics of privacy and utility for distorted matrices by average mutual information and average distortion rate in information-theoretic model with much more rigorous mathematical foundation.

3.3.1. Quantification for Privacy. Mutual information can be interpreted as the information preserved in the process of receiver from information source. Similarly, privacy can be defined as average mutual information in terms of discrete random probability space as in Scenario 4. Similar to noisy channel between symmetric discrete source and receiver, client-side and server-side can communicate by distorted matrices as shown in Table 1.

Our goal is to optimize distorted matrices by designed metrics inspired from noisy channel. In detail, privacy can be defined as average mutual information $\mathbf{I}(\mathbf{X}; \mathbf{Y})$ that the server-side obtained from client-side, namely, priori entropy

Table 1: Comparison of the two models.

Counterpart	Comparison
Source versus client-side	Discrete memoryless information source $\mathbf{X}$ Individual $U_i \in \mathbf{U}$ with categorical data $x_i \in \mathbf{X}$
Receiver versus server-side	Discrete information receiver $\mathbf{Y}$ Distorted version $y_i \in \mathbf{Y}$ Noisy channel
Channel versus distorted matrices	$\mathbf{M} = \begin{pmatrix} p_{x_1x_1} & \cdots & p_{x_1x_n} \\ \vdots & \ddots & \vdots \\ p_{x_nx_1} & \cdots & p_{x_nx_n} \end{pmatrix}$, sum of each row is 1

$\mathbf{H}(\mathbf{X})$ minus posterior entropy $\mathbf{H}(\mathbf{X} \mid \mathbf{Y})$. The value of privacy is equal to average uncertainty about $\mathbf{X}$ that remained in $\mathbf{Y}$

$$\begin{aligned} \mathbf{I}(\mathbf{X};\mathbf{Y}) &= \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X} \mid \mathbf{Y}) \\ &= \sum_{\mathbf{X}} P(x) \log \frac{1}{P(x)} \\ &\quad - \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{1}{P(x \mid y)} \\ &= \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{1}{P(x)} \\ &\quad - \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{1}{P(x \mid y)} \\ &= \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{P(x \mid y)}{P(x)} \\ &= \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{P(xy)}{P(x)P(y)} \\ &= \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{P(y \mid x)}{P(y)}. \end{aligned} \tag{23}$$

Undoubtedly, the upper bound of $\mathbf{I}(\mathbf{X};\mathbf{Y})$ can be restricted with obvious proof by

$$\mathbf{I}(\mathbf{X};\mathbf{Y}) \leq \min\left[\mathbf{H}(\mathbf{X}), \mathbf{H}(\mathbf{Y})\right]. \tag{24}$$

From the perspective of adversary, the smaller the value of $\mathbf{I}(\mathbf{X};\mathbf{Y})$ is, the better the privacy is protected. According to Jensen's inequality and (23), average mutual information is bounded by $0 \leq \mathbf{I}(\mathbf{X};\mathbf{Y}) \leq \min\{H(\mathbf{X}), H(\mathbf{Y})\}$ which is suitable for the bound of privacy. Metric functions of $\mathbf{f}$ in privacy computing can be represented by

$$\text{privacy} = \sum_{\mathbf{X},\mathbf{Y}} P(xy) \log \frac{P(y \mid x)}{P(y)}. \tag{25}$$

3.3.2. Quantification for Utility. In information-theoretic model of privacy computing, data utility can be quantified by average distortion rate which is the measure of how much $\mathbf{X}$ and $\mathbf{Y}$ are distorted. Distortion function and distortion matrix are represented by

$$d(\mathbf{X},\mathbf{Y}) \geq 0, \tag{26}$$

$$d(x_i, y_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j, \end{cases} \tag{27}$$

$$\mathbf{D} = \begin{bmatrix} d(a_1,b_1) & d(a_1,b_2) & \cdots & d(a_1,b_r) \\ d(a_2,b_1) & d(a_2,b_2) & \cdots & d(a_2,b_r) \\ \vdots & \vdots & & \vdots \\ d(a_r,b_1) & d(a_r,b_2) & \cdots & d(a_r,b_r) \end{bmatrix}. \tag{28}$$

Distortion function defined in (26) and (27) is nonnegative. The distortion matrix can be computed in (28). Finally, average distortion rate can be computed as

$$\begin{aligned} \overline{\mathbf{D}} &= \mathbf{E}\left[d(x,y)\right] = \sum_{\mathbf{X},\mathbf{Y}} P(xy)\, d(x,y) \\ &= \sum_{i=1} \sum_{j=1} P(y_i)\, P(x_j \mid y_i)\, d(x_i, y_j). \end{aligned} \tag{29}$$

Thus, utility is defined by average distortion rate; the smaller its value, the higher its utility. Metric functions of $\mathbf{g}$ in privacy computing can be represented by

$$\text{utility} = \sum_{i=1} \sum_{j=1} P(y_i)\, P(x_j \mid y_i)\, d(x_i, y_j). \tag{30}$$

Finally, our work is focused on minimizing the two objectives of privacy and utility simultaneously under specified constraints. In next section, we will illustrate our solution to the problem illustrated in (31) by our proposed discrete MOPSO

$$\begin{aligned} \mathbf{V} - \min \quad & \mathbf{F}(\mathbf{M}) = \left[\mathbf{f}(\mathbf{M}), \mathbf{g}(\mathbf{M})\right]^T \\ \text{s.t.} \quad & \text{discrete probabilty space} \\ & \varepsilon\text{-differential privacy} \\ & \det(\mathbf{M}) \neq 0. \end{aligned} \tag{31}$$

Notably, constraint that is determinant of $\mathbf{M}$, $\det(\mathbf{M})$, is nonzero which guarantees that $\mathbf{M}$ is invertible.

Input: Parameters of Np, Nr, Nm, $maxgen$, dim
Output: Final optimal set repository.
Step 1. Initialization:
Do
Step 1.1. Initialization in particles of position ***POS***, Velocity ***VEL***, ***PBEST***, ***GBEST***;
Step 1.2. Initialization in repository ***REP***;
Step 1.3. Initialization in mutation archive ***ARC***;
Step 1.4. Check constraints:
(1) Each column in ***POS*** is bounded by $\max_{x=1,\ldots,t}(\max_{x=1,\ldots,t} p_{xy}/\min_{x=1,\ldots,t} p_{xy}) \leq e^{\varepsilon}$;
(2) Sum of each column in ***POS*** is 1;
(3) Determinant of ***POS*** cannot be 0.
While constraints cannot be satisfied simultaneously, return to ***Step 1.1.***
Step 2. Repeat:
Step 2.1. Updating ***VEL*** and ***POS*** for each particle according to velocity updating mechanism and check constraints for new ***POS*** as ***Step 1.1***;
Step 2.2. Calculate fitness under evaluation function for each particle and update ***PBEST*** and ***GBEST***;
Step 2.3. Update ***REP*** by Pareto dominance and hyper-volume respectively;
Step 2.4. Mutation is operated on ***ARC*** while $Nm > 0$. Two strategies of mutation percentage is set to 1/3 and guided-random mechanism is a partial imitation of ***GBEST*** in ***REP*** under boundary constraints;
Step 3. Termination:
If stopping criterion is achieved, halt and check the combination of ***REP*** and ***ARC*** by Pareto dominance. Output the final optimal RR matricesfrom ***REP*** and ***ARC***;
Otherwise, loop to ***Step 2***.

Algorithm 1: Discrete MOPSO for optimizing for RR matrices.

4. MOPSO in Privacy Computing

In a simplified condition that RR matrices are generated from the range of $\{0, 1/d, 2/d, \ldots, 1\}$ (d is integer), the NP-hard problem of searching optimal RR matrices reaches an astronomical number of $\binom{d+n-1}{d}^n$, while the real search boundary of our scheme is the whole probability space which is much more complex than the condition above. Therefore, we provide discrete MOPSO to optimize conflicting goals of privacy and utility simultaneously and evaluate different RR schemes with fewer tunable variables as well as fast convergent rate.

4.1. Outline of Algorithm. The two objectives of privacy and utility are derived from RR matrices which can reflect the nature of privacy information in RR schemes. In detail, average mutual information and average distortion rate are constrained by ε-differential privacy and probability boundary. Our proposed discrete MOPSO is comprised of the following stages as shown in Algorithm 1. (1) Particles are initialized in a structure of discrete position, velocity, personal-best, global-best, repository, and mutation archive with dynamic hypervolume; (2) velocity updating is redefined by activation function of ReLU which can retain best genes from global-best for new discrete position; (3) repository updating is guaranteed by hypercube quality and diversity by hypervolume in combination with roulette wheel selection and crowding distances dynamically and adaptively; (4) performance of mutation archive is determined by two strategies of mutation percentage and guided-random mechanism to maintain diversity of computation; (5) constraints judgment: ε-differential privacy guarantees the rigorous protection bounds against arbitrary background knowledge. In each column of RR matrix, the ratio of maximum and minimum is bounded by privacy budget. The sum of each column is 1 and determinant of each matrix is nonzero; (6) termination judgment. Notably, the combination of repository and mutation archive are the final output of optimal RR matrices after Pareto dominance check. The process of our algorithm is illustrated in the following and parameters are listed in Abbreviations.

4.2. Initialization. RR matrix is coded as position of particle, **POS**, in which sum of each column is normalized to 1 and determinant of which is nonzero under constraint of ε-differential privacy. **POS_fit** is a vector of privacy and utility for each particle evaluated by two objective functions from (25) and (30). Initialization of velocity **VEL** is a randomized matrix with the same size as **POS**. Personal-best **PBEST** is initialized by **POS** and **POS_fit**. **GBEST** is selected from **PBEST** of each particle at random. Repository **REP** and mutation archive **ARC** are initialized as empty sets at first.

4.3. Velocity Updating. According to continual velocity updating rule in (32), we redefine the operation of velocity in a discrete form. Inspired by activation function of rectified linear unit (ReLU), velocity is discretized into a copy of certain features from **GBEST** by our scheme of ReLU in (32)

Input: **ARC** and mutation percentage **P**
Output: New particles.
Step 1. Randomized mutation
Select particles from **ARC** according to percentage **P**;
Each selected particle is re-randomized;
P $*$ *dim* columns in each particle are replaced by newly generated columns which still satisfy constraints as in Algorithm 1;
Calculate fitness for each newly generated particle.
Step 2. Guided mutation
Select the remaining particles from **ARC**;
Each particle is mutated by the guidance of **GBEST**;
(1 − **P**) $*$ *dim* columns in each particle are replaced by corresponding columns in **GBEST** which still satisfy constraints as in Algorithm 1;
Calculate fitness for each newly generated particle.

Algorithm 2: Mutation for discrete MOPSO.

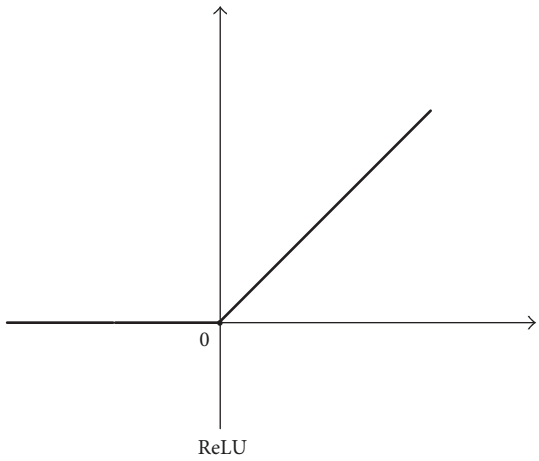


Figure 4: Activation function of ReLU.

and (33) as shown in Figure 4. Each element in **VEL** is a representation of certain excellent episodes from **GBEST**. In this paper, cognitive and social components of c_1, c_1 are fixed to value 2. Inertia weight of ω is set to 0.4. r_1 and r_2 are random factors generated from the interval of 0 and 1

$$\begin{aligned}\mathbf{VEL} = \mathrm{ReLU}\left(\omega * \mathbf{VEL} + c_1 * r_1 * (\mathbf{PBEST} - \mathbf{POS})\right.\\ \left. + c_2 * r_2 * (\mathbf{GBEST} - \mathbf{POS})\right)\end{aligned} \tag{32}$$

$$\mathrm{ReLU} = \max(0, x) = \begin{cases} 0, & \text{if } x \le 0 \\ x, & \text{if } x > 0. \end{cases} \tag{33}$$

In detail, **VEL** is transformed by ReLU by the threshold of 0. Utilizing the operator of AND with **GBEST** in (34), outstanding episode truncated from **GBEST** is preserved directly. Then corresponding position in **POS** is substituted by the outstanding episode maintained in **VEL** with replacing operator of $\oplus$. After normalization of **POS**, a new position is updated in (35)

$$\mathbf{VEL} = \mathbf{VEL} \text{ AND } \mathbf{GBSET} \tag{34}$$

$$\mathbf{POS} = \text{normalize } (\mathbf{VEL} \oplus \mathbf{POS}). \tag{35}$$

Overall, an illustrative example of velocity updating for each particle can be depicted in Figure 5.

4.4. Repository Updating. Repository **REP** is the archive of good RR matrices in the representative form of **POS**. In order to maintain diversity and optimize distribution, our updating mechanism is based on hypervolume and crowd distance. Hypervolume is divided by hyperlimits and hypercube dynamically and adaptively according to the boundaries of each Pareto solution in **REP**. Each hypercube bounded by hyperlimits is assigned with hyperquality based on the number of particles inside. **GBEST** is selected by roulette wheel in the hypercube with higher hyperquality. When particles exceed the capacity of repository, extra particles with the smallest crowding distances, which are measured by (36), are removed and transferred to the mutation archive **ARC** for the use of mutation

$$d(X_i) = \frac{1}{n-1}\sqrt{\sum_{\mathbf{Y}\in\mathbf{REP}-\mathbf{X}_i}^{dim} (X_i - \mathbf{Y})^2}, \tag{36}$$

where *dim* is the dimension of X_i and n denotes the number of particles in **REP**. In other words, crowding distance is a measure of average distance with all the other particles in **REP**.

4.5. Strategies for Mutation. To maintain diversity and enhance ability to escape from local optimum, two strategies for mutation are designed: fixed mutation percentage and guided-randomization as shown in Algorithm 2.

4.6. Boundaries Checking. Position of all particles are bounded by ε-differential privacy when

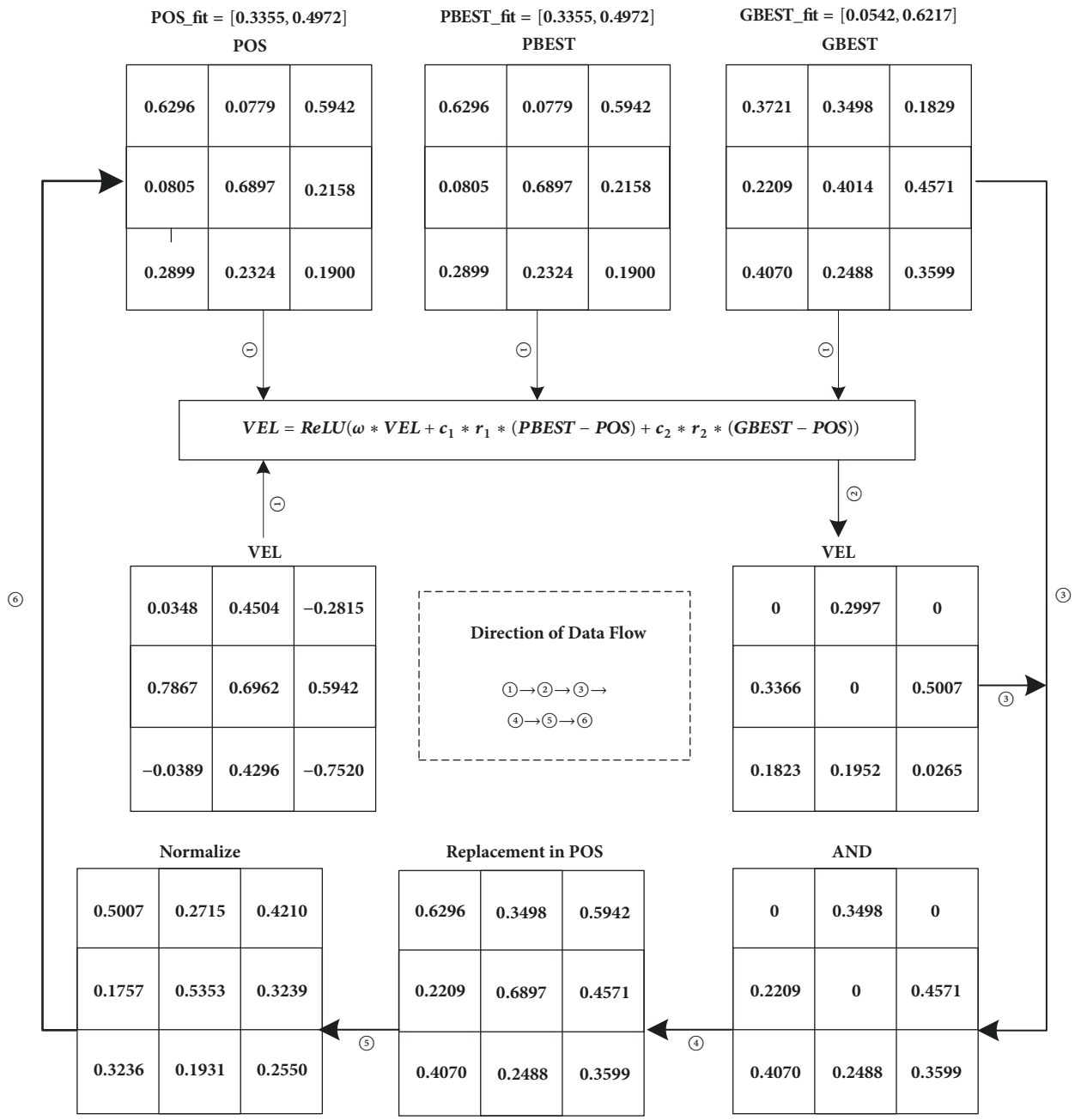


Figure 5: Example of velocity updating.

$\max_{x=1,\ldots,t}(\max_{x=1,\ldots,t} p_{xy}/\min_{x=1,\ldots,t} p_{xy}) \leq e^{\varepsilon}$, which means that the maximum in each column is no more than e^{ε} times of the minimum. The tunable parameter of privacy budget ε is regulated to satisfy different users' privacy-preserved requirements. In addition, mutation is performed in the phase of particle updating when certain position is out of the bound.

4.7. Termination Checking. Termination criterion for MOPSO can be set as follows: (1) limiting the maximum iteration number; (2) among certain number of consecutive generations, quality of repository **REP** does not improve. When termination criterion is met, the combination set of optimal RR matrices in Pareto dominance from **REP** and **ARC** is output as the final solution. In our work for termination, we combine both methods for the reason that either criterion can guarantee a diversified optimal solution set for different requirements and can benefit from the fast convergent feature of MOPSO.

4.8. Complexity Analysis

4.8.1. Space Complexity. In our scheme of discrete MOPSO, two RR matrices set, namely, **REP** and **ARC,** are needed, both

Table 2: Parameters of our algorithm.

Notation	Value
Population size Np	100
Repository size Nr	200
Mutation archive size Nm	200
Maximum number of iterations $Maxgen$	200
Mutation percentage p	1/3
Initial weight w	0.4
Cognitive coefficients c_1, c_2	2
Number of grids in **REP** $Ngrid$	20
hypercube_limits in **REP** will change dynamically as with boundaries of particles.	

of which occupy complexity of $\mathbf{O}((Nr+Nm)*dim^2)$. Nr and Nm are the size of **REP** and **ARC** with particles' dimension of dim. All particles need a complexity of $\mathbf{O}(Np*dim^2)$, where Np is the size of population. Thus, the total space complexity of discrete MOPSO for RR matrices is $\mathbf{O}((Np+Nm+Nr)*dim^2)$.

4.8.2. Time Complexity. The main time complexity is occupied by Pareto dominance check and repository updating in Algorithm 1. For ease of representation, let n be the number of solution. The evaluation number is $2C_n^2$, namely, $\mathbf{O}(n^2)$. Initialization in **Step 1** needs $\mathbf{O}(Np*n^2)$ operations of Pareto dominance check. Repository updating in **Step 2** occupies complexity of $\mathbf{O}(n^2)$. Thus, the worst running time complexity of our scheme is $\mathbf{O}(\max(Np, maxgen)*n^2)$, where $maxgen$ is the iteration number.

5. Experiments and Discussion

We conduct our experiments on two kinds of datasets: synthesis datasets and real datasets from UCI. Performance of our proposed discrete MOPSO is compared with Wanner, FRAPP, UP, and SPEA2 in [34]. Parameters of discrete MOPSO are shown in Table 2.

5.1. Pareto Front. RR matrices that are generated from Warner, UP, and FRAPP have been proved to obtain the same Pareto front distribution [34]. Thus, for ease of illustration, we take FRAPP as a comparison. Synthesis datasets are generated in the following way: diagonal element of p in RR matrices is generated from [0, 1] with the stride of 0.001. 999 RR matrices can be obtained. Random numbers are chosen from different distribution in a single-dimension set of 10000 records with 8 different categories. Nondominated solutions that are beyond differential privacy are deleted in advance. For different distributions, we set privacy budget $\boldsymbol{\varepsilon} = 2$. In addition, when privacy budget ε varies, the results are featured with similar trends. Six kinds of Pareto fronts of different RR schemes are presented in Figure 6. Corresponding statistics from boxplot are shown in Figure 7.

In Figure 6, we can conclude that metric of utility obtained by our approach can reach a wider range than FRAPP. Our solutions are almost in dominance of all those in FRAPP, which shows the better global optimization capability by discrete MOPSO. In terms of privacy, we can find more Pareto solutions for RR matrices than FRAPP. FRAPP just consider RR matrices by a single metric of accuracy, while our method can optimize two metrics of privacy and utility simultaneously.

We take Gamma distribution in Figure 6(f) as an example. Only 389 out of 999 RR matrices in FRAPP satisfy differential privacy when $\varepsilon = 2$, while our approach has the ability to maintain a wider range of privacy and utility by adopting **REP** and **ARC**. Mutation and velocity updating operation enhance the searching ability especially in a wider range of privacy. On the other hand, it can be calculated that $\mathbf{H}(\mathbf{X})$ is 2.6741 and $\mathbf{H}(\mathbf{Y})$ = 2.9358. Maximum of privacy falls within the scope of minimum of $\mathbf{H}(\mathbf{X})$ and $\mathbf{H}(\mathbf{Y})$ which can prove the conclusion in Section 3.3.1.

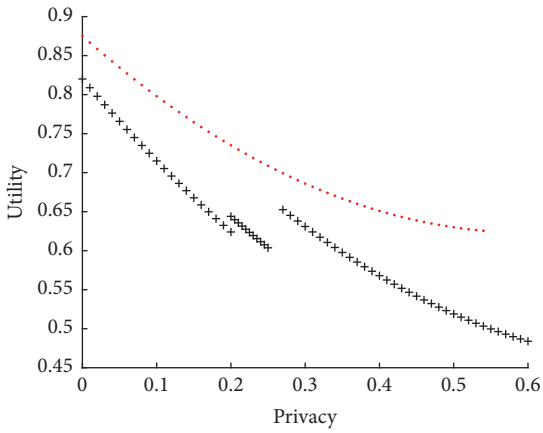
In Figure 7, it obviously shows that RR matrices optimized by our method have a much wider range of privacy and utility than FRAPP, especially in the metric of privacy by average mutual information. It can be summarized that lower distortion rate responds to better utility and better performance of privacy is in accordance with less average mutual information.

5.2. Privacy Budget. Privacy budget is the most important index of balancing the trade-off between privacy and utility. Under the bound of differential privacy that $\max_{u=1,\ldots,t}(\max_{v=1,\ldots,t} p_{uv}/\min_{v=1,\ldots,t} p_{uv}) \le e^{\varepsilon}$, effect of different privacy budgets ($\varepsilon = 0, 0.1, 0.5, 1, 1.5, 2$) on the number of RR matrices is shown in Figure 8.

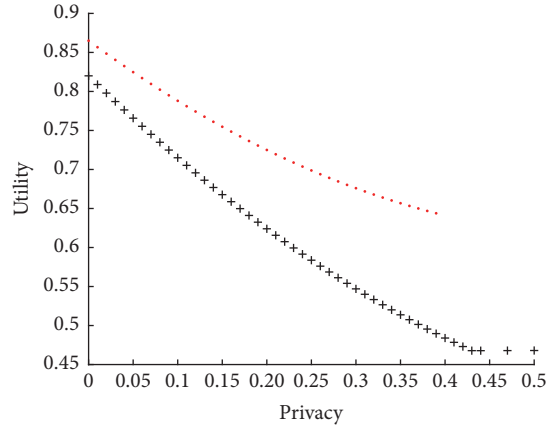
In Figure 8, the case of $\varepsilon = 0$ means that each element in RR matrix is equal, which is the special condition of maximum privacy with minimum utility according to information-theoretic model. Notably, when $\varepsilon = 0, 0.1, 0.5, 1$, the size of **REP** and **ARC** is set to 200 and when $\varepsilon = 1.5$, 2, the size is set to 400 for evaluating the further exploitative ability of our method in Figure 8(a). In Figure 8(b), where $\varepsilon = 0, 0.1, 0.5$, size of **REP** and **ARC** is set to 200. It can be concluded that our method has a better performance on finding more RR matrices. As ε grows larger, FRAPP can find more RR matrices. However, our method is focused on a more diversified distribution of Pareto solution with much higher quality on the trade-off between privacy and utility. As with decrease in ε, privacy is strongly preserved. There is a sharp drop in the number of RR matrices obtained by FRAPP, while our method keeps almost stable due to the adoption of **REP** and **ARC**.

5.3. Case Study. We adopt two datasets from UCI: (1) car evaluation data set, including 1728 instances of 4 classes of uncc. (70.023%), acc. (22.222%), good (3.993%), and v. good (3.762%) with 6 attributes; (2) the game of Connect-4 opening database containing all legal 8-ply positions from 67557 instances and 42 attributes in 3 classes of win (65.83%), loss (24.62%), and draw (9.55%).

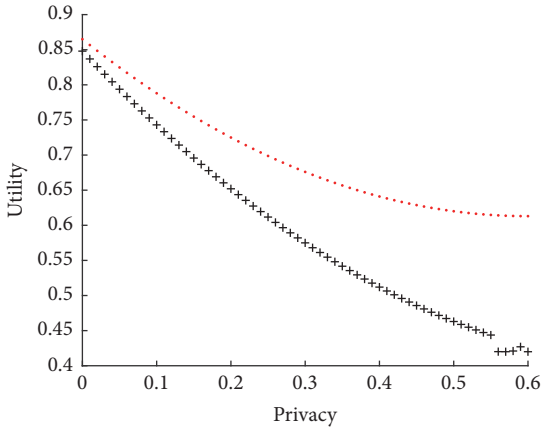
In Figure 9, under a strong bound of privacy budget $\varepsilon = 0.5$, we can conclude that our method can obtain a wider range of Pareto front in terms of privacy and utility than



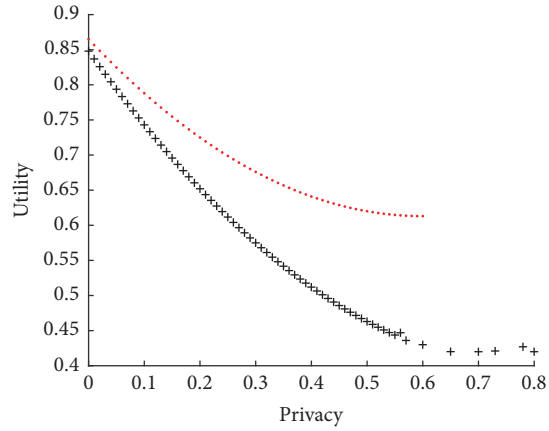
(a) Uniform distribution by linear standard uniform transformation



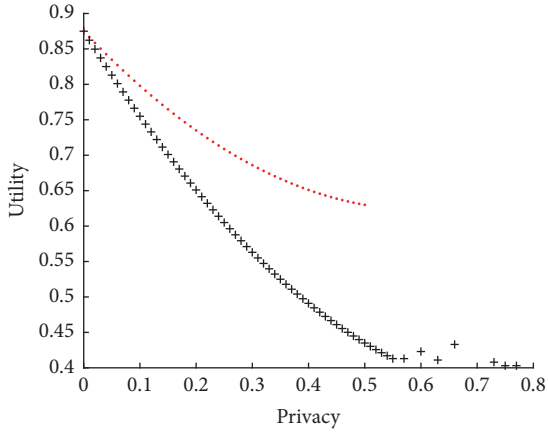
(b) Student's t distribution with 3 degrees of freedom



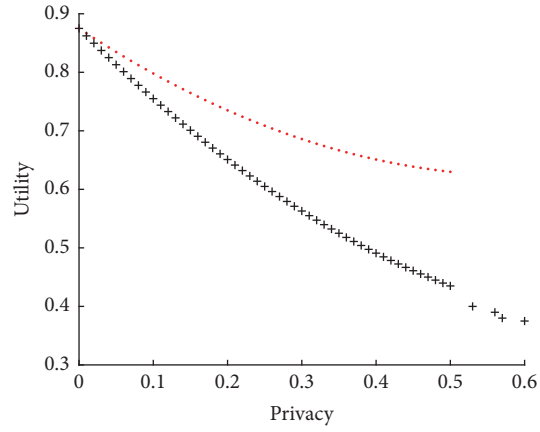
(c) Chi-square distribution with 8 degrees of freedom



(d) Binomial distribution ($N = 10000$, $p = 0.3$)

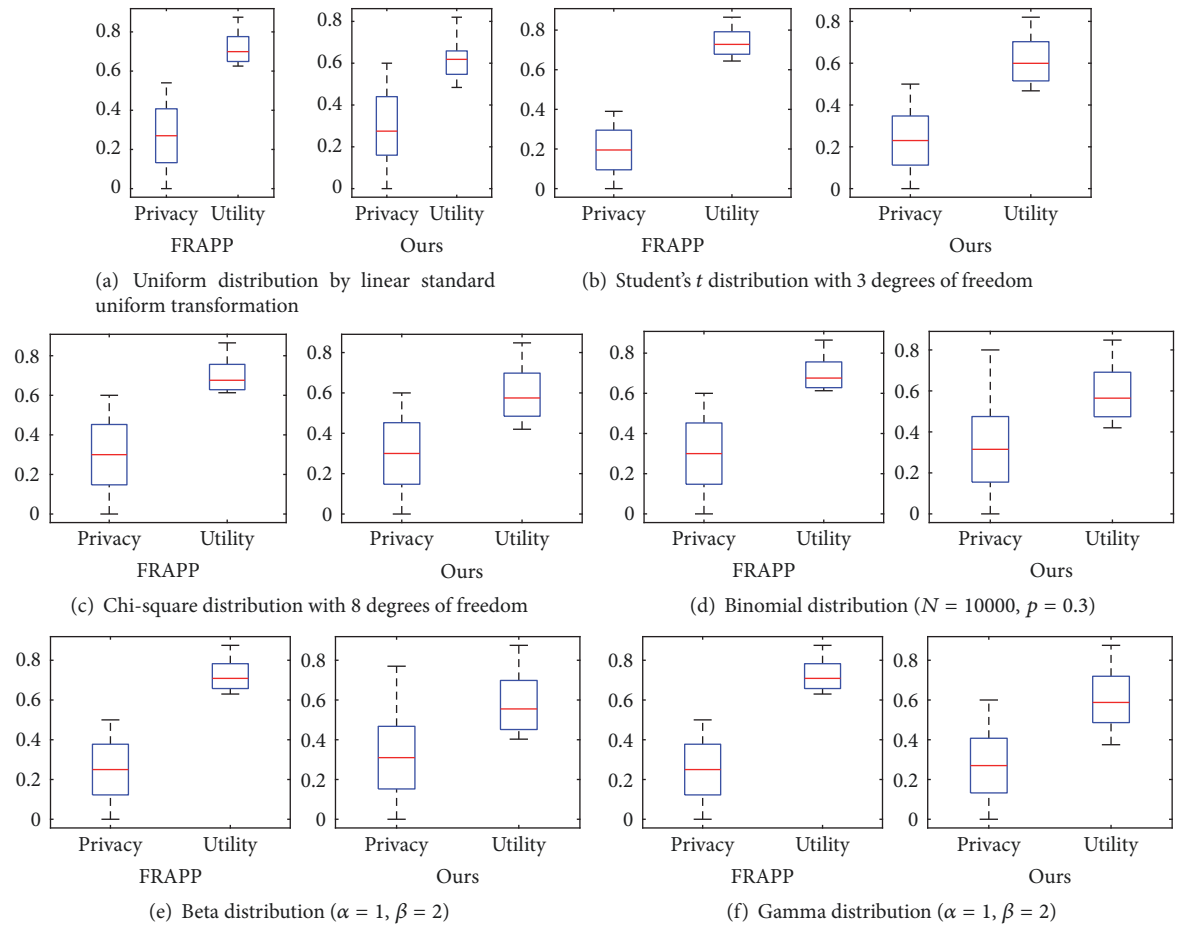


(e) Beta distribution ($\alpha = 1$, $\beta = 2$)



(f) Gamma distribution ($\alpha = 1$, $\beta = 2$)

Figure 6: Pareto front of different distribution with privacy budget $\varepsilon = 2$.



(a) Uniform distribution by linear standard uniform transformation

(b) Student's t distribution with 3 degrees of freedom

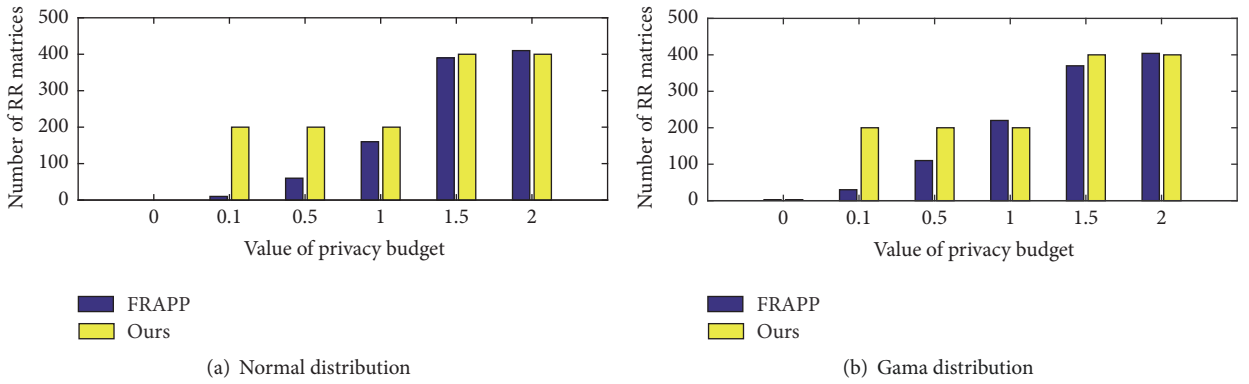
(c) Chi-square distribution with 8 degrees of freedom

(d) Binomial distribution ($N = 10000$, $p = 0.3$)

(e) Beta distribution ($\alpha = 1$, $\beta = 2$)

(f) Gamma distribution ($\alpha = 1$, $\beta = 2$)

FIGURE 7: Boxplot of different distribution with privacy budget $\varepsilon = 2$.



(a) Normal distribution

(b) Gama distribution

FIGURE 8: Effect on number of RR matrices with different ε.

FRAPP for car evaluation, especially in the metric of privacy. For Connect-4, both methods have nearly the same range and statistics in boxplot figure. Meanwhile, our approach can still gain better statistics in Pareto dominance of FRAPP as shown in Figures 9(a) and 9(c). In addition, when $\varepsilon = 2$, we can estimate the best utility of original data distribution by the two methods. Our method has a better performance than FRAPP as to the metric of utility as illustrated in Figure 10.

In addition, to fully test our method, we compared it with OptRR [34] in the aspect of running time for obtaining 200 RR matrices in **REP**. Dimension of RR matrices, D, is set from 5 to 150. OptRR is developed from evolutionary multiobjective optimization method, namely, SPEA2. The experiment is conducted by Intel core i5-4590 *CPU@3.30* GHz, 4 GB RAM, and coded in matlab on Windows 7-64 bit. Results of running time are shown in Table 3.

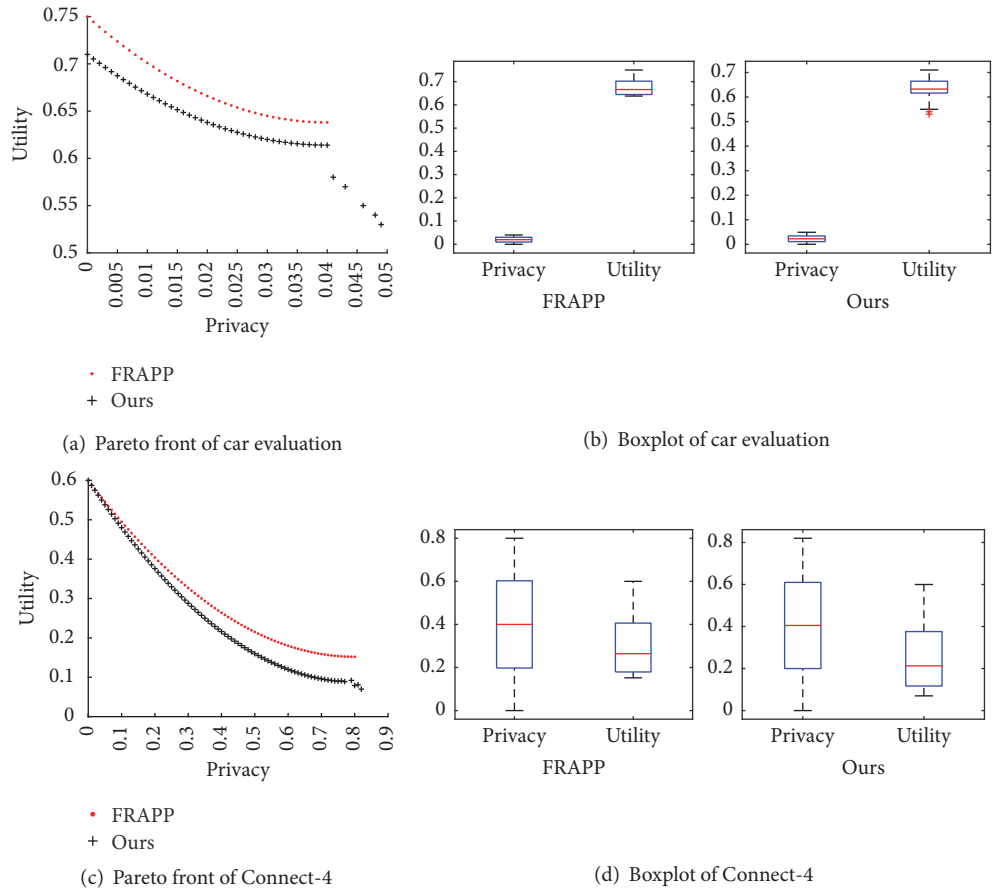


Figure 9: Pareto front and boxplot with $\varepsilon = 0.5$.

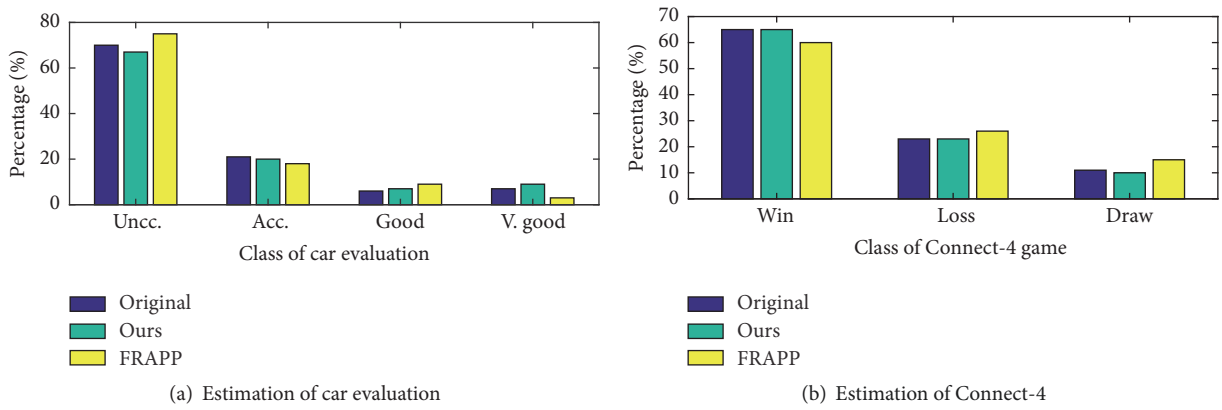


Figure 10: Estimation of datasets with $\varepsilon = 2$.

The running time results are obtained from the mean value of 20 independent experiments in Table 3. Our approach has a faster convergence rate than OptRR in finding optimal RR matrices, especially when dimension of RR matrices grows. Larger dimensions of RR matrices challenge computing capability as well as the optimization strategies in MOPSO and OptRR. The simple operators in our methods like AND and Replacement are beneficial in reducing complexity and,also, the advantage of MOPSO in fast convergence plays a key role in speeding up the searching process.

6. Conclusion

In this paper, we formulate the information-theoretic model as to quantification for privacy and establish an open framework of privacy computing. In terms of optimizing RR matrices in information collection by randomized

Table 3: Running time results (seconds).

Approach	Dimension of RR matrices					
	$D = 5$	$D = 10$	$D = 20$	$D = 50$	$D = 100$	$D = 150$
Ours	6.57	6.87	7.08	10.88	27.12	53.28
OptRR	20.79	67.83	140.73	301.35	697.26	1762.43

response, we derived metrics for privacy and utility, namely, average mutual information and average distortion rate. Furthermore, our proposed approach is the first solution to optimized RR matrices using discrete MOPSO under information-theoretic model. Our discrete MOPSO is proposed to solve the double-objective optimization problem under differential privacy. Two mutation strategies in extra mutation archive and velocity updating mechanism are successful in helping MOPSO escape from local optimum. The experimental results on synthesis datasets and real datasets from UCI show that our approach outperforms existing state-of-the-art works in the aspects of Pareto front, privacy budget, and data distribution reconstruction, especially excellent in privacy protection. Moreover, we hope to contribute our optimal Pareto solutions of extensive RR schemes to the public for the future study. Some extensive work can be applied to other domains of privacy computing.

Abbreviations

Illustration of General Channel Model

$P(\mathbf{X})$: Probability that source takes value $\mathbf{X} \in \{a_1, a_2, \ldots, a_s\}$
$P(\mathbf{Y})$: Probability that receiver takes value $\mathbf{Y} \in \{b_1, b_2, \ldots, b_s\}$
$P(\mathbf{Y} \mid \mathbf{X})$: Probability that receiver takes value $\mathbf{Y}$ when source value $\mathbf{X}$ is presented.

Tuples of Privacy Computing

X: Private information
P: Participants involved in privacy computing
M: Quantifying metrics on private attributes
Φ: Constraints on privacy disclosure
Ω: Operation on private information
T: Effective privacy computing tools and paradigms
Δ: Supplementary set of privacy computing system.

Description of Parameters in Our Approach

Np: Population size
Dim: Dimension
Nr: Repository size
Nm: Mutation archive size
maxgen: Maximum number of iterations.

Conflicts of Interest

The authors declare that there will not be any conflicts of interest regarding the publication of this manuscript.

Acknowledgments

This work was supported by the Chinese National Natural Science Foundation (Grant no. U1603261) and the Natural Science Foundation of Xinjiang Province, China (Grant no. 2016D01A080). Also, the authors would like to thank their partners at their research lab for their generous gifts in supporting this research. Genuine thanks are, especially, due to Professors Yi Qu, Yiliang Han, Anming Gong, and Zhen Liu for their beneficial advice.

References

[1] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2014.

[2] Y. Guo, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding," *Neurocomputing*, vol. 187, no. C, pp. 27–48, 2016.

[3] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.

[4] Y. Liu, B. Guo, H. Du, Z. W. Yu, D. Zhang, and C. Chen, "FooD-Net: Optimized On Demand Take-out Food Delivery using Spatial Crowdsourcing," in *Proceedings of the MobiCom'17*, Association for Computing Machinery, Snowbird, UT, USA, 2017.

[5] M. S. Farash and M. A. Attari, *A Provably Secure And Efficient Authentication Scheme for Access Control in Mobile Pay-TV Systems*, Kluwer Academic Publishers, Norwell, MA, USA, 2016.

[6] https://search-id.com/aol/about.

[7] S. Hansell, "AOL removes search data on vast group of web users," *New York Times*, 2006.

[8] https://www.netflixprize.com/faq.html.

[9] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *Computer Science*, 2006.

[10] http://www.epic.org/privacy/inrefacebook/EPIC-FacebookComplaint.pdf.

[11] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing NYC Taxi Data: Does It Matter?" in *IEEE International Conference on Data Science and Advanced Analytics*, pp. 140–148, Montréal, Canada, 2016.

[12] https://www.thestar.com/business/2017/09/07/equifax-says-data-breach-may-affect-143-million-people-in-us.html.

[13] L. Sweeney, *K -Anonymity: A Model for Protecting Privacy*, vol. 10, World Scientific Publishing, Singapore, 2002.

[14] 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, 2007.

[15] 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*, pp. 106–115, IEEE, Istanbul, Turkey, 2007.

[16] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 754–759, ACM, Philadelphia, PA, USA, 2006.

[17] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," in *In Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 689–700, ACM, Beijing, China, 2007.

[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Theory of Cryptography Conference*, vol. 3876, pp. 265–284, Springer, New York, NY, USA, 2006.

[19] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[20] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *In Proceedings of the 42nd ACM symposium on Theory of computing*, pp. 715–724, ACM, Cambridge, MA, USA, 2010.

[21] X. Li, J. Yang, Z. Sun, and J. Zhang, "Differential privacy for edge weights in social networks," *Security & Communication Networks*, vol. 2017, no. 4, 10 pages, 2017.

[22] T. Gao, F. Li, Y. Chen, and X. K. Zou, "Preserving local differential privacy in online social networks," in *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications*, vol. 10251, Springer, Guilin, China, 2007.

[23] C. Lin, Z. Song, Q. Liu, W. Sun, and G. Wu, "Protecting Privacy for Big Data in Body Sensor Networks: A Differential Privacy Approach," in *in the proceedings of International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 163–172, Springer, Beijing, China, 2015.

[24] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," 2016, https://arxiv.org/abs/1606.05053v1.

[25] B. Mishra, S. White, D. Hosseini et al., "Storage, Retrieval, Analysis, Pricing, and Marketing of Personal Health Care Data Using Social Networks, Expert Networks, and Markets," WO, WO/2013/188838, 2013.

[26] F. Kargl, A. Friedman, and R. Boreli, "Differential privacy in intelligent transportation systems," in *ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pp. 107–112, ACM, Budapest, Hungary, 2013.

[27] I. Roy, S. T. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and privacy for MapReduce," in *Proceedings of the Symposium on Networked Systems Design and Implementation (NSDI)*, vol. 10, pp. 297–312, USENIX – Advanced Computing Systems Association, Lombard, IL, USA, 2010.

[28] Vijay Srinivas Agneeswaran, *Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives*, Pearson Education, London, UK, 2014.

[29] R. Khatri and A. Kumar, "Spark (lightning-fast cluster computing) application intelecommunication sector to prevent customer churn out," *International Journal of Computer & Mathematical Sciences*, vol. 4, pp. 124–126, 2015.

[30] W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 25, pp. 505–510, DBLP, Washington, Dc, USA, 2003.

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

[32] M. G. Gong, Q. Cai, X. W. Chen, and L. J. Ma, "Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 82–97, 2014.

[33] Y. Wang, X. T. Wu, and D. H. Hu, "Using Randomized Response for Differential Privacy Preserving Data Collection," *EDBT/ICDT Workshops*, 2016.

[34] Z. Huang and W. Du, "OptRR: Optimizing randomized response schemes for privacy-preserving data mining," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE'08*, pp. 705–714, mex, April 2008.

[35] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 247–255, Santa Barbara, California, USA, May 2001.

[36] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 439–450, 2000.

[37] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in *ACM SIGMOD International Conference on Management of Data*, vol. 23, pp. 251–262, DBLP, Baltimore, Maryland, USA, 2005.

[38] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proceedings of the 23rd ACM Conference on Computer and Communications Security, CCS 2016*, pp. 192–203, aut, October 2016.

[39] S. Agrawal and J. R. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining," in *International Conference on Data Engineering, 2005. ICDE, 2005*, vol. 18, pp. 193–204, IEEE, Tokyo, Japan, 2005.

[40] L. Ninghui, W. Qardaji, and D. Su, "Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy," pp. 32-33, 2010, Corr abs/1101.2604.

[41] F. H. Li, Y. Jia, and Li. H. Jia, "Privacy computing: concept, connotation and its research trend," *Journal on Communications*, vol. 37, pp. 1–11, 2016.

[42] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the International Symposium on Information Theory*, vol. 1, pp. 267–281, Akadémiai Kiadó, Budapest, Hungary, 1973.

[43] A. Gilbert and A. McMillan, "Local Differential Privacy for Physical Sensor Data and Sparse Recovery," https://arxiv.org/abs/1706.05916.

[44] T. Kulkarni, G. Cormode, and D. Srivastava, "Marginal Release Under Local Differential Privacy," https://arxiv.org/abs/1711.02952.

[45] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–66, 1965.

[46] B. Kang, J. Wang, and D. Shao, "Attack on Privacy-Preserving Public Auditing Schemes for Cloud Storage," *Mathematical Problems in Engineering*, vol. 2017, Article ID 8062182, 2017.

[47] Z. Lin, X. Xiao, Y. Sun, Y. Zhang, and Y. Ma, "A Privacy-Preserving Intelligent Medical Diagnosis System Based on Oblivious Keyword Search," *Mathematical Problems in Engineering*, vol. 2017, Article ID 8632183, 7 pages, 2017.