Construction of Differentially Private Empirical Distributions from a low-order Marginals Set through Solving Linear Equations with l_2 Regularization

Evercita C. Eugenio[1] and Fang Liu[2]

[1] Sandia National Laboratories, Livermore, CA 94550, USA
eceugen@sandia.gov
[2] University of Notre Dame, Notre Dame, IN 46556, USA
fang.liu.131@nd.edu

Abstract. We introduce a new algorithm, **C**onstruction of d**I**fferentially **P**rivate Empirical Distributions from a low-order marginals set t**H**rough solving linear **E**quations with l_2 **R**egularization (CIPHER), that produces differentially private empirical joint distributions from a set of low-order marginals. CIPHER is conceptually simple and requires no more than decomposing joint probabilities via basic probability rules to construct a linear equation set and subsequently solving the equations. Compared to the full-dimensional histogram (FDH) sanitization, CIPHER has drastically lower requirements on computational storage and memory, which is practically attractive especially considering that the high-order signals preserved by the FDH sanitization are likely just sample randomness and rarely of interest. Our experiments demonstrate that CIPHER outperforms the multiplicative weighting exponential mechanism in preserving original information and has similar or superior cost-normalized utility to FDH sanitization at the same privacy budget.

Keywords: differentially private empirical distributions and synthetic data, sign and statistical significance (SSS), full-dimensional histogram low-order marginals, computational storage and memory

1 Introduction

1.1 Background and Motivation

When releasing data sets for research and public use, protection of individual private information while still maintaining good utility of the data is of extreme importance. Even if a data set is de-identified, a data intruder may still be able to identify subjects by linking to other publicly available information [1–5].

[1] The research was funded by the US National Science Foundation Awards #1546373 and #1717417. The publication has been assigned the Sandia National Laboratories identifier SAND2021-0088 C.

The intensified concerns on privacy call for more rigorous and mathematically sound privacy protection concepts and frameworks when sharing information. Differential privacy (DP) [6, 7] has emerged as one of the most powerful concepts to achieve that goal. DP provides rigorous mathematical guarantee for privacy protection without making strong assumptions about the intruder's background knowledge. One of the applications of DP is to generate differentially private distributions and individual-level synthetic or surrogate data so that data users may perform analysis on their own as if they had the original data. A simple way to sanitize data with minimal distributional assumptions on the original data is to sanitize the full-dimensional histogram (FDH), an empirical estimator of the joint distribution of all the attributes in the data. The approach is simple but has drawbacks when the data are multi-dimensional. First, there are likely a lot of empty cells in the histogram when its dimensionality p is relatively large. Second, the high-order interactions among the attributes implicitly preserved by the FDH hardly represent any meaningful population-level signals. Lastly, it can be computationally costly to store or sanitize the FDH for a large p.

1.2 Our Contributions

We propose a novel procedure, namely, **C**onstruction of d**I**fferentially **P**rivate Empirical Distributions from a low-order marginals set t**H**rough solving linear **E**quations with l_2 **R**egularization (CIPHER), to generate differentially private empirical distributions from which individual-level data surrogate or synthetic data can be easily obtained. CIPHER is based on the general knowledge that the population-level signals in real-life data are oftentimes contained in a set of low-order marginals. The advantages of CIPHER and its practical significance's are summarized as follow.

- CIPHER is conceptually simple and requires nothing than decomposing joint probabilities among attributes via basic probability rules to construct and solve a linear equation set. It does not impose strong assumptions on the local data. The computational cost for solving the equation set is expected to be low once the equation set is constructed.
- CIPHER can automatically correct the inconsistency across the marginals of the same variable that appear in multiple histograms caused by differentially private sanitization, without explicitly incorporating constraints.
- Compared to the FDH sanitization, the set of low-order marginals that CIPHER employs has drastically lower requirements for computer storage and memory. For example, compared to 9,765,625 cells resultant from the FDH of 10 attributes with 5 levels each, there is a 95.4% and 99.99% reduction in the number of cells – 62,200 and 8,440, respectively – if CIPHER uses the set of four-way histograms (210) and the set of two-way histograms (45), respectively.
- If a data user is provided with a set of differentially private low-order marginals with inconsistent counts due to sanitization, she may apply CIPHER to generate individual-level data to meet her analysis needs, without incurring additional privacy costs.

1.3 Related Work

There exist some methods that generate differentially private empirical distributions or synthetic data from a set of low-order statistics, even though they might be proposed originally for different purposes. Each approach has its own pros and cons, and differs from CIPHER in either the formulation of the query sets, or computationally, or methodologically.

Barak et al [8] propose a method based on Fourier transforms. The linear programming employed by the method is a bottleneck for this algorithm especially for large p. Chen et al [9] form differentially private histograms from attribute clusters. The formation of optimal attribute clusters is an NP-hard problem and the authors introduce an approximation algorithm that does not guarantee optimality. Compared to these two methods, CIPHER is less costly computationally, though it requires a careful layout of the equation set from which the sanitized empirical distribution is calculated. Liu [10] proposes a model-based approach to generate differentially private synthetic data (modips) in the Bayesian framework. The modips has practical limitations in terms of both model construction as well as sufficient statistics sanitization for certain data types or large p. Machanavajjhala et al [11] demonstrate that the Multinomial-Dirichlet model sanitization leads to poor inferences due to data sparsity when it is applied to release the commuting patterns of the US population data. Bowen and Liu [12] also show that the approach has worse performance than the FDH sanitization via the Laplace mechanism and the modips approach at the same privacy budget. CIPHER, compared to the modips and Multinomial-Dirichlet model methods, does not require specification of a statistical model on the original data. PrivBayes [13] has some similarity with CIPHER in the sense that both rely on a reduced set of relationships among the attributes to capture the signals in the data. On the other hand, PrivBayes is different from CIPHER in that it explores the conditional independence among the attributes to approximate the joint distribution of the attributes via a directed acyclic graph – Bayesian network, whereas CIPHER does not need data-driven selection of the set of queries or formulation of a model, and can save all privacy budget toward the sanitization process. The price paid by CIPHER is that the set of queries it generates the empirical distribution from might not be the most representative of the underlying population signals and relations among the attributes compared to the Bayesian network selected by PrivBayes by spending a certain amount of privacy cost. Hardt et al [14] propose the iterative Multiplicative Weights via Exponential Mechanism (MWEM) approach. The MWEM algorithm achieves a near optimal bound on the l_∞ error for the queries used to generate synthetic samples if the number of iterations T is optimized. However, it can be very challenging to choose the optimal T and the accuracy of MWEM is highly dependent on T. MWEM is an iterative procedure and each iteration incurs privacy cost due to it accessing the original data to fetch the query selected by the Exponential mechanism, which is subsequently sanitized by the Laplace mechanism. CIPHER is non-iterative and has different privacy cost than MWEM. We will discuss more on the differences between CIPHER and MWEM, and

between CIPHER and PrivBayes in Section 2.2 after presenting the detailed steps of the CIPHER algorithm.

The rest of the paper is organized as follows. Section 2 introduces the CIPHER procedure. Section 3 compares the CIPHER with several other sanitization methods on the statistical utility of the synthetic data in simulated and real-life data. It also proposes the SSS (Sign and Statistical Significance) assessment to evaluate the inferences based on differentially private synthetic data against the original inferences. Section 4 provides some concluding remarks and discusses future research directions.

2 CIPHER

2.1 Preliminaries

Consider a data set D and a query or a set of queries $\mathbf{f}$ about D. DP provides a rigorous mathematical conceptual framework to protect individual privacy information when releasing the query results to $\mathbf{f}$.

Definition 1 (ϵ-differential privacy [6]) *A randomized mechanism $\mathcal{R}$ satisfies ϵ-differential privacy if for all data sets D_1 and D_2 differing on one individual and all result subsets S to query $\mathbf{f}$, $e^{-\epsilon} \leq \frac{\Pr[\mathcal{R}(\mathbf{f}(D_1))\in S]}{\Pr[\mathcal{R}(\mathbf{f}(D_2))\in S]} \leq e^{\epsilon}$ for $\epsilon > 0$.*

The formulation of privacy via the DP is robust and guards against the worst-case scenario as it does not impose any assumptions about the behavior or the background knowledge of data intruders. D_1 and D_2 differing by one individual can be interpreted in two ways: D_1 is one individual more or less than D_2, or D_1 and D_2 are of the same size but have difference in attributes values in exactly one individual. ϵ is often referred as the privacy budget and is pre-specified. The smaller ϵ is, the more privacy protection is imposed on the individuals in the data. Interested readers may refer Bowen and Liu [12] and Liu et al [15] for brief discussions on the choice of ϵ. Besides the pure ϵ-DP in Definition 1, there are several relaxed versions, such as the (ϵ, δ)-DP and (ϵ, δ)-probabilistic DP. The former relaxes the bounds on the ratio of $\Pr(\mathcal{R}(\mathbf{f}(D_1) \in S)/\Pr(\mathcal{R}(\mathbf{f}(D_2)) \in S))$, whereas the latter bounds the probability the pure ϵ-DP is violated. In both cases, setting δ at 0 reduces to the pure ϵ-DP.

In practice, a data set can be queried multiple times. Every time a query result is released, some privacy is lost for the individuals in the data set. The sequential composition and parallel composition principles presented in Definition 2 are useful for tracking and counting of privacy costs when designing differentially private mechanisms and releasing query results form a data set.

Definition 2 (sequential composition and parallel composition [16]) *Let f_i for $i = 1, ..., K$ represent a set of queries on data D and ϵ be the total privacy budget. Denote by $\mathcal{R}_i$ a randomization mechanism of ϵ_i-DP. The sequential composition states that the sequence of $\mathcal{R}_i(D)$ provides $(\sum_i \epsilon_i)$-DP. The parallel composition states that the sequence of $\mathcal{R}_i(D_i)$ provides $\max_i \epsilon_i$-DP, where $\{D\}_{i=1,...,K}$ are arbitrary disjoint subsets of D.*

There are a variety of mechanisms to provide differentially private results. Interested readers may refer to Dwork and Roth [17] for some of the commonly used DP mechanisms. Here we mention the Laplace mechanism, that is used in the experiments in Section 3.

Definition 3 (Laplace mechanism [6]) *The Laplace mechanism of ϵ-DP generates sanitized query results as in $\mathbf{f}^*(D) \sim Lap(\mathbf{f}(D), \Delta_\mathbf{f}/\epsilon)$, where $\Delta_\mathbf{f} = \max_{D_1,D_2} \|\mathbf{f}(D_1) - \mathbf{f}(D_2)\|_1$ is the l_1 global sensitivity of query $\mathbf{f}$, for all data sets D_1, D_2 differing by one element.*

The larger $\Delta_\mathbf{f}$ is, the more noise would be injected to $\mathbf{f}$ to satisfy the ϵ-DP. Generalization of the Laplace mechanism include the Gaussian mechanism and Generalized Gaussian mechanism that is built upon the l_p norm ($p \geq 1$) [17, 18], among others. Definition 4 presents the exponential mechanism, which is used in the MWEM procedure implemented in the experiments in Section 3.

Definition 4 (exponential mechanism [19]) *Let u be a utility function that assigns a score to each possible output of a query to data D. The Exponential mechanism that satisfies ϵ-DP releases query result $f^*(D)$ with probability*

$$\exp(u(f^*(D); D)\tfrac{\epsilon}{2\delta_u}) / \int u(f^*(D); D)\tfrac{\epsilon}{2\delta_u} d(f^*(D)),$$

where δ_u is the maximum change in score u with one element change in data D.

2.2 The CIPHER Procedure

We propose the CIPHER procedure to generate differentially private empirical distributions and individual-level synthetic data from a set of low-order marginals. As mentioned in Section1, the main motivation for CIPHER is to reduce the dimension of the query set from which individual-level data can be generated to save on storage and memory, while preserving the population-level signals in the original data. Figure 1 shows the drastic reduction in the number of

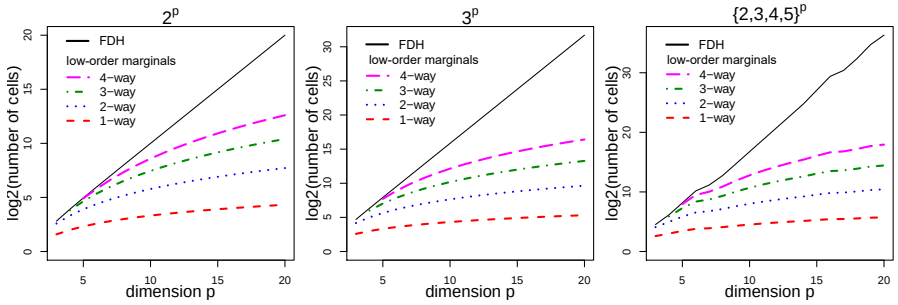

Fig. 1. log(number of stored cell in low-order marginals) vs the number of attributes p (leftmost: each attribute is binary; middle: each attributes has 3 levels; rightmost: the number of levels ranges from 2 to 5 among the p attributes)

cells that need to be stored if the sets of 1-way, 2-way, 3-way, and 4-way low-order marginals are used in place of the full table for varying p (the number

of attributes in the original data). The degree of the low-order marginals from which higher-order marginals are generated is allowed to grow with p; but it should always be kept in mind that interactions of very high orders are rarely of interest in real life and are also hard to explain.

The CIPHER algorithm is presented in Algorithm 1. In brief, the algorithm starts with a set of low-order marginals $\mathcal{Q}$ and arrives at a solution of the differentially private empirical distribution via a stepwise but non-iterative fashion, without involving complex sampling algorithms. The low-order marginals in $\mathcal{Q}$, which do not have to be of the same dimension, capture the important signals and relationships among the attributes in the original data. Two special cases of

Algorithm 1: CIPHER

input : original data D $(n \times p)$; query set $\mathcal{Q}$; privacy budget ϵ; number of empirical distributions or synthetic data sets m; l_2 regularization constant λ.

output: differentially private empirical distributions $P(\mathbf{X})^{(l)}$ or data sets $\tilde{D}^{(l)}$ for $l = 1, \ldots, m$.

Denote the lowest dimension of the marginals in $\mathcal{Q}$ by p_0;
for $l = 1, \ldots, m$ **do**
  Sanitize all queries $\in \mathcal{Q}$ via a mechanism of ϵ-DP (e.g., $\tilde{q}_i^{(l)} = q_i + \text{Lap}(0, \epsilon/(m|\mathcal{Q}|))$ for $i = 1, \ldots, |\mathcal{Q}|$);
  for $j = p_0 + 1, \ldots, p$ **do**
    List all j-way marginals $\mathcal{T}_j$;
    for *each* $q_i \notin (\mathcal{T}_{j+1} \cap \mathcal{Q})$ **do**
      1) Denote the variables involved in query q_i by $\mathcal{X}_i$ and let $p_i = |\mathcal{X}_i|$;
      2) Randomly pick a variable out of $\mathcal{X}_i$ and label it as X_{i1}, and the rest as $X_{i2}, \ldots, X_{i,p_i}$. Denote the number of bins or levels of X_{ik} by K_{ik} for $k = 1, \ldots, p_i$;
      3) **for** $k = 2, \ldots, (p_i - 1)$ **do**
        Define $\mathbf{b}_k = P(X_{i1} \neq K_{i1} | \mathbf{X}_i \setminus (X_{i1}, X_{ik})$ $= \sum_{X_{ik}} P(X_{i1} \neq K_{i1}, X_{ik} | \mathbf{X}_i \setminus (X_{i1}, X_{ik})) = \mathbf{A_k z_k} =$ $\sum_{X_{ik}} P(X_{ik} | \mathbf{X}_i \setminus (X_{i1}, X_{ik})) P(X_{i0} \neq K_{i1} | \mathbf{X}_i \setminus (X_{i1}, X_{ik}), X_{ik})$, where $\mathbf{z}_k$ is the conditional probability of $(X_{i1} \neq K_{i1})$ given the rest of variables in $\mathcal{X}_i$, $\mathbf{A}_k$ is either observed or calculated from step $j - 1$, and $(X_{i1} \neq K_{i1})$ represents the vector $(X_{i1} = 1, \ldots, X_{i1} = K_{i1} - 1)$.;
      end
      4) Let $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_{p_i - 1})^T$, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{b}_{p_i - 1})^T$, and $\mathbf{A} = \text{Diag}\{\mathbf{A}_1, \ldots, \mathbf{A}_{p_i - 1}\}$; solve for $\mathbf{z}$ from $\mathbf{Az} = \mathbf{b}$ as in $\mathbf{z} = (\mathbf{A^T A} + \lambda \mathbf{I})^{-1} \mathbf{A^T b}$, where I is the identity matrix;
      5) Calculate the private empirical joint probability the variables in $\mathcal{X}_i$: $P(\mathcal{X}_i) = \mathbf{z} \cdot P(\mathbf{X}_i \setminus X_{i1})$;
    end
  end
  Correct negativity and normalize the private empirical joint probability $P(\mathbf{X})^{(l)} = P(X_1, ..., X_p)^{(l)}$, and generate private data $\tilde{D}^{(l)}$ of size n from $P(\mathbf{X})^{(l)}$ if needed;
end

$\mathcal{Q}$ are the single p-way full table and the set of p one-way contingency tables, respectively. Forming $\mathcal{Q}$ can be guided by the domain knowledge so not to consume the additional privacy in the original data. If the domain knowledge is not available or the data curator prefers to choose $\mathcal{Q}$ using the information of the original data, the total privacy budget will need to be divided between the selection of $\mathcal{Q}$ and the CIPHER algorithm itself. In the rest of the discussion, we assume $\mathcal{Q}$ is preset before the application of the CIPHER algorithm.

Claim 1 *The CIPHER algorithm satisfies ϵ-DP.*

The satisfaction of DP in CIPHER is straightforward to establish. The only time at which the original data are probed during the application of CIPHER is when the queries in $\mathcal{Q}$ are sanitized. All together, the data are accessed mK times with a privacy budget of $\epsilon/(mK)$ per access. Per the sequential composition, the total privacy budget for releasing the privacy-preserving empirical distribution is maintained at $(mK)\epsilon/(mK) = \epsilon$.

We recommend setting the number of sanitized distribution sets m in the algorithm at a small number, say 1 to 5. $m > 1$ is specified when there is a need to account for the randomness and uncertainty in the released information due the sanitization and synthesis. In addition, as long as m is not too large so that the total privacy budget is not spread too thin over the multiple sets (each synthetic set receives $1/m$ of the total privacy budget per the sequential composition theorem), the precision gained by averaging over m sets of synthetic data could outweigh the additional noises introduced from releasing multiple sets than a single set. In the case of statistical inferences based on m sets of synthetic data are of interest, they can be obtained through the inferential combination rules in [10].

The reason for using the l_2 regularization (aka the Tikhonov regularization) to solve for $\mathbf{z}$ from $\mathbf{Az} = \mathbf{b}$ is that the columns of $\mathbf{A}$ are linearly dependent and A^TA is not of full rank. The l_2 regularization is known for solving ill-posed problems like $\mathbf{Az} = \mathbf{b}$ when the solution $\mathbf{z}$ is not unique due to the singularity of $\mathbf{A}$ [20, 21]. It adds a small positive constant λ to $\mathbf{A^TA}$ and calculates $\mathbf{z} = (\mathbf{A^TA} + \lambda\mathbf{I})^{-1}\mathbf{A^Tb}$. Since $\mathbf{A}$ is block-diagonal, taking the inverse of $\mathbf{A^TA} + \lambda\mathbf{I}$ is computationally cheap, relatively speaking, even if the linear equation set is large. Regarding the choice of hyper-parameter λ, our empirical studies suggest that the solutions are relatively insensitive to λ in the case of CIPHER as long as λ is relatively small ($o(1)$).

If one or more attributes end up appearing in ≥ 2 marginals in $\mathcal{Q}$, then after the sanitization, the empirical distributions of the shared attributes would be inconsistent across different low-order marginals. The good news that the inconsistency is automatically averaged out for CIPHER when solving the linear equation set, without a need for ad-hoc step to correct for the inconsistency, because the formulation of the linear equation set implicitly builds in the constrains. This offers a great advantage over the methods that employ ad-hoc correction procedures (e.g.,[8, 22]).

The cell probabilities in the low-order marginals in $\mathcal{Q}$ after DP sanitization can be < 0 or ≥ 1, so are the solutions for the conditional probabilities in some intermediate steps of the CIPHER algorithm. We suggest performing a one-time correction in the last step of generating the empirical distribution.

Both CIPHER and MWEM may use a pre-specified set of linear queries to generate differentially private empirical distributions, but they differ methodologically and algorithmically. First, MWEM relies on an iterative multiplicative weighting procedure whereas CIPHER is not iterative. Second, the queries in CIPHER are sanitized one time through a DP mechanism (say the Laplace sanitizer) before being fed to the algorithm. By contrast, each iteration in the MWEM algorithm incurs a privacy cost due to it accessing the original data to fetch the query selected by the exponential mechanism, which is subsequently sanitized by the Laplace mechanism. As a result, the two algorithms spend different amounts of privacy cost per query for a given total privacy budget. Suppose the total budget is ϵ, and the number of queries in $\mathcal{Q}$ is denoted by $|\mathcal{Q}|$. If we use equal allocation of the privacy budget, then each query in $\mathcal{Q}$ gets a budget of $\epsilon/|\mathcal{Q}|$ in the CIPHER algorithm. The sanitization of each query selected by the exponential mechanism costs $\epsilon/(2T)$ in the MWEM algorithm. On the other hand, a query can be selected multiple times throughout the T iterations. Let c_k denote the number of times that $q_k \in \mathcal{Q}$ is selected among the T iterations. Note $\sum_{k=1}^{|\mathcal{Q}|} c_k = T$. Unless $c_k/(2T) > |\mathcal{Q}|^{-1}$ or $c_k/\sum_{k=1}^{|\mathcal{Q}|} c_k > 2|\mathcal{Q}|^{-1}$, then the budget allocated to query q_k in the MWEM algorithm would always be smaller than that in CIPHER. In other words, the selection probability for a query needs to at least double the average selection probability $(1/|\mathcal{Q}|)$ in order to that query to receive more privacy budget in MWEW than the amount of budget it receives in CIPHER. In addition, our own experiences from running the MWEM algorithm suggest that choosing the "right" number of iterations T for MWEM can be challenging. T too small is not sufficient to allow the empirical distribution to fully capture the signals summarized in the queries; and T too large would lead to a large amount of noises being injected as the privacy budget has to be distributed across the T iterations, eventually leading to a useless synthetic data set as each iteration costs privacy. PrivBayes offers similar benefits on data storage and memory, similar to CIPHER, given that PrivBayes is built in the framework of Bayesian network that is is known for its ability of saving considerable amounts of memory over full-dimensional tables if the dependencies in the joint distribution are sparse. On the other hand, PriBayes starts with model building that costs privacy budget. It is also well known approximate structure learning of a Bayesian network is NP-complete. In addition, Bayesian networks would force attributes in a data set to be in a causal relationship. Finally, PrivBayes proposes a surrogate function for mutual information, on which the quality of the released data replies, requires some effort for efficient computation. In comparison, the underlying analytical and computational techniques for CIPHER are standard and require nothing than joint probability decomposition and solving linear equations.

2.3 Example: CIPHER for the 3-variable Case

We illustrate the CIPHER procedure with a simple example. Say the original data contain 3 variables ($p = 3$). Denote the 3 variables by V_1, V_2, V_3 with K_1, K_2 and K_3 levels, respectively. Let $\mathcal{Q} = \{\mathcal{T}(V_1, V_2), \mathcal{T}(V_2, V_3), \mathcal{T}(V_1, V_3)\}$ that contains all the 2-way contingency tables. Therefore, $p_0 = 2$ in Algorithm 1. WLOG, suppose V_3 is X_0 in Algorithm 1. We first write down the relationships among the probabilities, which are

$$\begin{cases} \Pr(V_3|V_1) = \sum_{V_2} \Pr(V_3, V_2|V_1) = \sum_{V_2} \Pr(V_3|V_1, V_2) \Pr(V_2|V_1) \\ \Pr(V_3|V_2) = \sum_{V_1} \Pr(V_3, V_1|V_2) = \sum_{V_1} \Pr(V_3|V_1, V_2) \Pr(V_1|V_2) \end{cases}.$$

We now convert the above relationships into the equation set $\mathbf{b} = \mathbf{A}\mathbf{z}$. Specifically, $\mathbf{b} = (\Pr(V_3|V_1) \setminus \Pr(V_3 = K_3|V_1)$, and $\Pr(V_3|V_2) \setminus \Pr(V_3 = K_3|V_1))^T$ is a known vector of dimension $(K_1 + K_2)(K_3 - 1)$, $\mathbf{z} = \Pr(V_3|V_1, V_2) \setminus \Pr(V_3 = K_3|V_1, V_2)$ is of dimension $K_1K_2(K_3 - 1)$, $\mathbf{A}$ is a known diagonal matrix with $K_3 - 1$ identical blocks, and each block is a $(K_1 + K_2) \times (K_1K_2)$ matrix comprising the coefficients (i.e., $\Pr(V_1|V_2), \Pr(V_2|V_1)$ or 0) associated with $\mathbf{z}$. After $\mathbf{z}$ is solved from $\mathbf{b} = \mathbf{A}\mathbf{z}$, the joint distribution of $\Pr(V_1, V_2, V_3)$ is calculated by $\mathbf{z} \cdot \Pr(V_1, V_2)$. The experiments in Section 3 contain more complicated applications of CIPHER.

3 Experiments

We run experiments with simulated and real-life data to evaluate CIPHER, and benchmark its performance against MWEM and the FDH sanitization in this paper. Both MWEM and FDH inspired our work, the former conceptually as stated in Section 2.2, while the latter motivated us to develop a solution with decreased storage cost for sanitized queries. In addition, more complex algorithms are unlikely to beat a simpler and easier-to-deploy flat algorithm such as FDH, per conclusions from previous studies [12, 23] when n, ϵ, or p is large. In addition, both MWEM and FDH are straightforward to program and implement. Our goal is to demonstrate in the experiments that CIPHER performs better than MWEM in terms of the utility of sanitized empirical distributions and synthetic data, and delivers non-inferior performance compared to FDH with significant decreased storage costs.

When comparing the utility of synthetic data generated by different procedures, we not only examine the degree to which the original information is preserved on descriptive statistics such as mean and l_q ($q > 0$) distance, we also examine the information preservation in statistical inferences on population parameters. Toward that end, we propose the *SSS assessment*. The first S refers to the Sign of a parameter estimate, and the second and third S' refer to the Statistical Significance of the estimate against the null value in hypothesis testing. Whether the sign and statistical significance in the estimate between the original and synthetic data are consistent leads to 7 possible scenarios (Table 1). Between the best and the worst scenarios, there are 5 other possibilities. II+ and I+ indicate an increase in Type II (false negative) and Type I (false positive) error

rates, respectively, from the original to the sanitized inferences, so do II- and I-, but the latter two also involve a sign change from the original to the sanitized inferences.

Table 1. Preservation of **S**igns and **S**tatistical **S**ignificance on an estimated parameter (the SSS assessment)

parameter estimate	Best		Neutral	II+	I+	II-	I-	Worst
matching **S**igns between non-private and sanitized?	Y	Y	N	Y	Y	N	N	N
non-private **S**tatistical **S**ignificance?	Y	N	N	Y	N	Y	N	Y
sanitized **S**tatistical **S**ignificance?	Y	N	N	N	Y	N	Y	Y

3.1 Experiment 1: Simulated Data

The data were simulated via a sequence of multinomial logistic regression models with four categorical variables and two samples size scenarios at $n = 200$ and $n = 500$, respectively. For the FDH sanitization, there are 36 ($2 \times 2 \times 3 \times 3$) cells in the 4-way marginals. For the CIPHER and MWEM algorithms, we consider 3 different query sets $\mathcal{Q}$: (1) $\mathcal{Q}_3$ contains all 3-way marginals, leading to 32 cells (88.9% of the 4-way); (2) $\mathcal{Q}_2$ contains all six 2-way marginals, leading to 20 cells (55.6% of the 4-way). Five privacy budget scenarios $\epsilon = (e^{-2}, e^{-1}, 1, e^1, e^2)$ were examined. To account for the uncertainty of the sanitization and synthesis in the subsequent statistical inferences, $m = 5$ synthetic data sets were generated. We run 1,000 repetitions for each n and ϵ scenario to investigate the stability of each method. When examining the utility of the differentially private data, we present the cost-normalized metrics wherever it makes sense. The cost is defined as the number of cells used to generate a differentially private empirical distribution (36 for FDH; 32 for CIPHER 2-way and MWEM 2-way; and 20 for CIPHER 3-way and MWEM 3-way).

In the first analysis, the average total variation distance (TVD) between the original and synthetic data sets was calculated for the 3-way, 2-way and 1-way marginals, respectively. Figure 2 presents the results. After the cost normalization, CIPHER 2-way performs the best overall, especially for small ϵ, and delivers similar performances as the FDH sanitization at large ϵ. CIPHER 3-way is inferior to CIPHER 2-way. There is minimal change in the performance of MWEM produces across ϵ.

In the second analysis, we examine the the l_∞ error for $\mathcal{Q}_2$ and $\mathcal{Q}_3$, the results of which are given in Figure 3 for at $n = 200$ (the findings are similar at $n = 500$ and are available in the supplementary materials). CIPHER 2-way and the FDH sanitizations are similar with the former slightly better at $\epsilon = e^{-2}$. CIPHER 3-way is second-tier behind CIPHER 2-way and the FDH sanitization. The performance of MWEM does not seem to live up to the claim that it yields the optimal l_∞ error for the set of queries that are fed to the algorithm [14]. This could be due to the fact that T, which is not an easy hyperparameter to tune, was not optimized in a precise way (though roughly using independent data) in our implementation of MWEM.

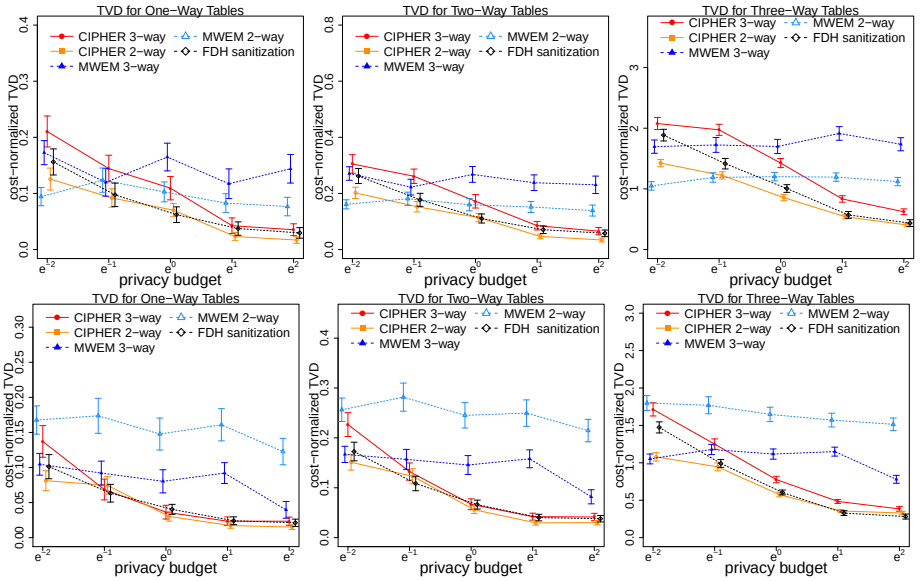

Fig. 2. Cost-normalized total variation distance (mean ± SD over 1,000 repeats) (top: $n = 200$; bottom: $n = 500$)

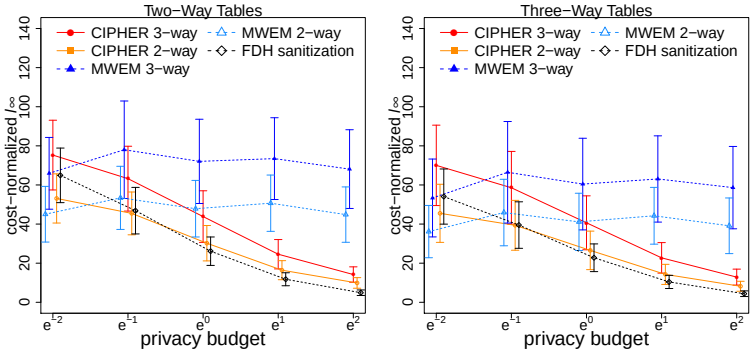

Fig. 3. Cost-normalized l_∞ (mean ± SD over 1,000 repeats) at $n = 200$

In the third analysis, we fitted the multinomial logit model with a binary attribute as the outcome and the others as predictors. The inferences from the $m = 5$ synthetic data sets were combined using the rule in [10]. The bias, root mean square error (RMSE), coverage probability (CP) of the 95% confidence interval (CI) were determined for each regression coefficient in the model. We present the results at $n = 200$ and $\epsilon = e^{-2}, 1, e^2$ in Figure 4 and those at $n = 500$ and for e^{-1} and e at $n = 200$ are listed in the supplementary materials. The observations are $n = 500$ are consistent with $n = 200$, and those at $\epsilon = e^{-1}$ and e when $n = 200$ are in between e^{-2} and 1, and between 1 and e^2, respectively in Figure 4. The 10 parameters from the model are listed on the x-axis. There is not much difference among CIPHER 2-way, CIPHER 3-way, and the FDH sanitization in cost-normalized bias or CP, but CIPHER 2-way delivers better performance in terms of cost-normalized RMSE (smaller) at all ϵ

compared to CIPHER 3-way, and at small ϵ compared to the FDH sanitization. CIPHER and the FDH sanitization deliver near-nominal CP (95%) across all the examined ϵ and both n scenarios while MWEM suffers severe under-coverage on some parameters especially at small ϵ. MWEM has the smallest cost-normalized RMSE for $\epsilon \leq 1$; but the RMSE values for CIPHER and the FDH sanitization catch up quickly and approach the original values as ϵ increases.

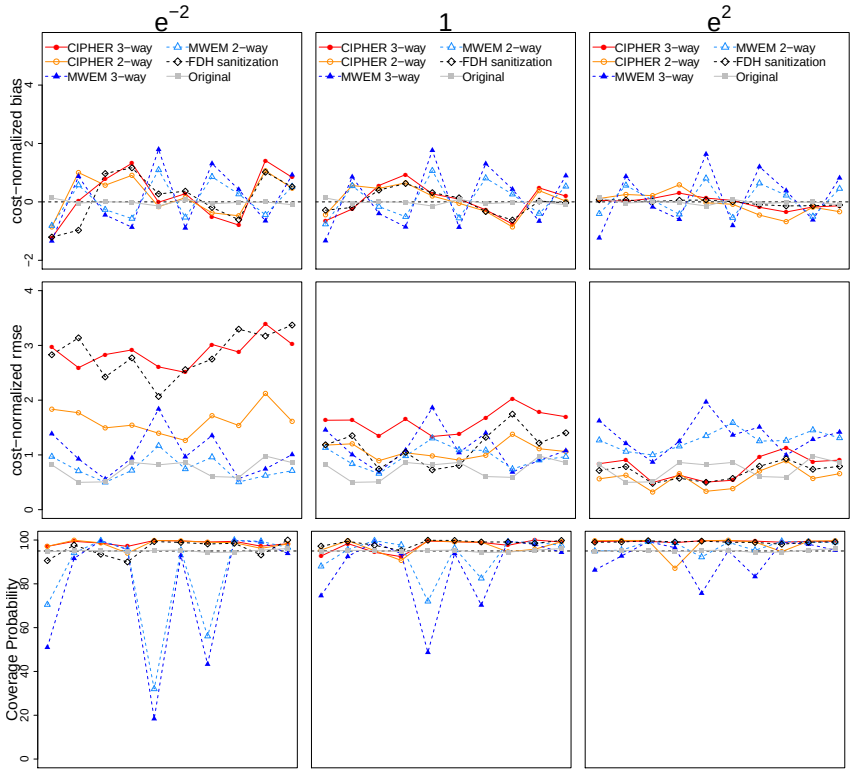

Fig. 4. Cost-normalized bias and root mean square error (rmse), and coverage probability (un-normalized) at $n = 200$

The results for the SSS assessment on the regression coefficients from the logistic regression are provided in Figure 5, un-normalized for the storage cost. A method with the longest red bar (best-case scenario as defined in Table 1) and the shortest purple bar (the worst-case scenario) would be preferable. The two inflated type II error (i.e., decreased power types) (the II+/orange bar and the I-/green bar) and neural (the gray bar) are acceptable, and the two inflated type I error types (I+/yellow bar and I-/blue bar) would preferably be of low probability. Per the listed criteria above, first, it is comforting to see the undesirable cases (purple+blue bars) are the shortest among all the 7 scenarios for each method. Second, the inferences improve quickly for CIPHER and the

FDH sanitization and rather slowly for MWEM as ϵ increases. Third, the FDH sanitization is the best performer in preserving SSS, especially for the medium valued ϵ, followed closely by CIPHER. Finally, even for CIPHER and the full table sanitization, there are always non-ignorable proportions of II+ (and II- when ϵ is small) especially when ϵ is as large as e^2, suggesting the sanitization decreases the efficiency of the statistical inferences, which is the expected price paid for privacy protection.

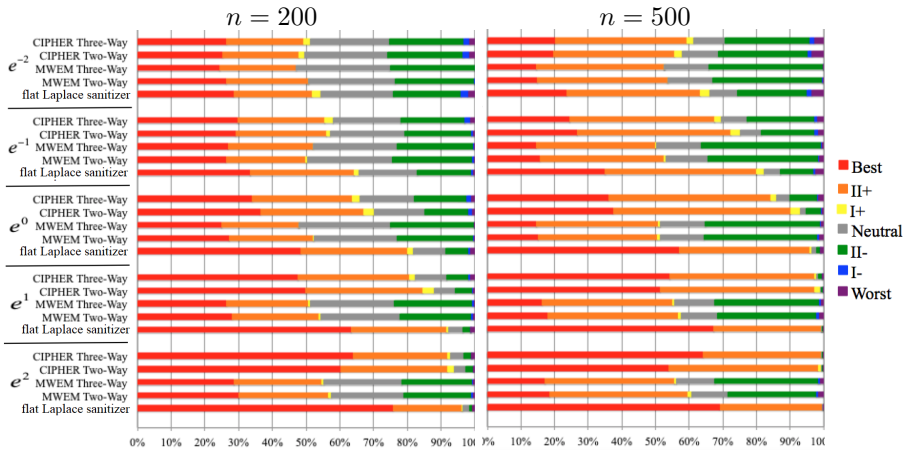

Fig. 5. The SSS (**S**igns and **S**tatistical **S**ignificance) assessment on the estimated regression coefficients for $n = 200$ and $n = 500$, un-normalized for storage cost

The cost-normalized log-odds of sanitized parameter estimates falling in the "best" category are provided in Figure 6 on the model parameters. The larger the odds, the more consistent the sanitized and the original inferences are on these parameters. Overall, the odds are similar for CIPHER 3-way, CIPHER 2-way, and the FDH sanitization, with CIPHER 3-way being the best at small ϵ. MWEM performs similarly as the other methods at small ϵ, but does not improve as ϵ increases.

3.2 Experiment 2: Qualitative Bankruptcy Data

The experiments runs on a real-life qualitative bankruptcy data set. The data were collected to help identify the qualitative risk factors associated with bankruptcy and is downloadable from the UCI Machine Learning repository [24]. The data set contains $n = 250$ businesses and 7 variables. Though the data set does not contain any identifiers, sensitive information (such as credibility or bankruptcy status) can still be disclosed using the pseudo-identifiers left in the data (such as industrial risk level or competitiveness level), or be used to be linked to other public data to trigger other types of information disclosure. The supplementary materials provides a listing of the attributes in the data.

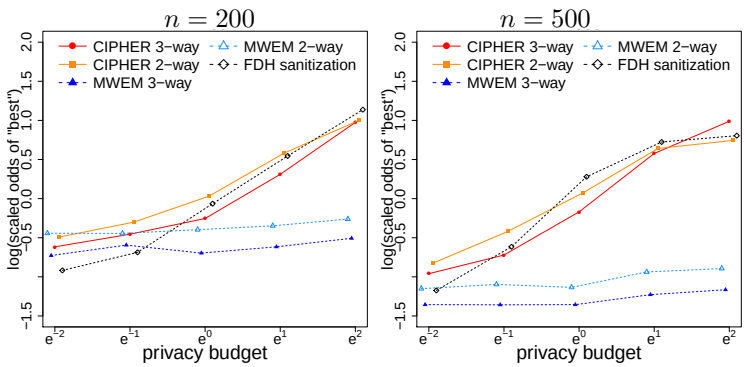

Fig. 6. Cost-normalized log(odds of the "best" category) in the SSS assessment on the estimated regression coefficients

$\mathcal{Q}$ employed by the CIPHER and MWEM procedures contains one 4-way marginal, six 3-way marginals, and three 2-way marginals, that were selected based on the domain knowledge, and computational and analytical considerations when solving the linear equations in CIPHER, without referring to the actual values in the data. More details are provided in the supplementary materials on how $\mathcal{Q}$ was chosen. The size of $\mathcal{Q}$ (the number of cell counts) is 149, which is about 10% of the number of cells counts (1,458 cells) in the FDH sanitization.

On the synthetic data generated by the three procedures, we ran a logistic regression model with "Class" as the outcome variable (bankruptcy vs non-bankruptcy) and examined its relationship with the other 6 qualitative categorical predictors [25]. We applied the SSS assessment to the estimated parameters from the logistic regression and the results are presented in Figure 7. The figure

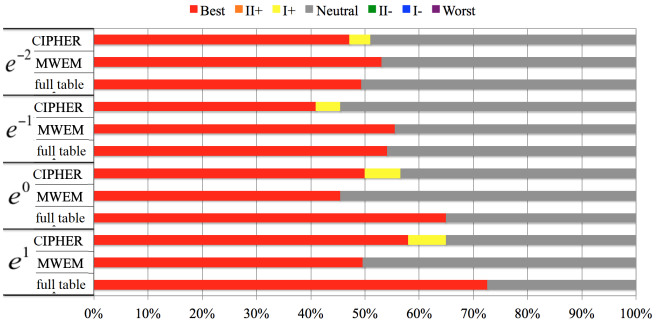

Fig. 7. The SSS assessment on the logistic regression coefficients in Experiment 2

suggests that all three methods perform well in the sense that the probability that they produced a "bad" estimate (the worst, II-, and I- categories) is close to 0, and the estimates are mostly likely to land in the "best" or the "neutral" categories. The FDH sanitization has the largest chance to produce estimates in

the "best" category for $\epsilon \geq e^{-1}$, at a much higher storage cost (~ 8 folds higher) than CIPHER and MWEM. MWEM has slightly better chance (50%) landing in the "best" category when $\epsilon \leq e^{-1}$ but does not improve for as ϵ increases.

We also performed an SVM analysis to predict "Class" given the other attributes on a testing data (a random set of 50 cases from the original). Per Table 2, CIPHER is the obvious winner at $\epsilon \leq 1$ with significantly better prediction accuracy than the other two and the accuracy is roughly constant. FDH is better than CIPHER at $\epsilon > e$, but with a ~ 8 -fold increase in the storage cost. The prediction accuracy remains $\sim 50\%$ for MWEM across all examined ϵ values, basically not much better than a random guess on the outcome.

Table 2. Prediction Accuracy (%) on "Class" via SVM in Experiment 2

ϵ	CIPHER	MWEM	FDH sanitization
e^{-2}	**67.8**	50.0	41.1
e^{-1}	**64.7**	51.3	55.5
1	**68.5**	51.0	63.8
e	77.8	47.2	**85.7**
e^2	90.3	47.3	**98.8**

4 Discussion

We propose the CIPHER procedure to generate differentially private empirical distributions from a set of low-order marginals. Once the empirical distributions are obtained, individual-level synthetic data can be generated. The experiment results implies that CIPHER delivers similar or superior performances to the FDH sanitization, especially at low privacy cost after taking into account the storage cost. CIPHER in general delivers significantly better results on all the examined metrics than MWEM. Both the CIPHER and MWEM procedures have multiple sources of errors in addition the sanitation randomness injected to ensure DP. For CIPHER, it is the shrinkage bias brought by the l_2 regularization; and for MWEM , it is the numerical errors introduced through the iterative procedure with a hard-to-choose T. The asymptotic version of both CIPHER and MWEM is the FDH sanitization when the low-order marginals set contains only one query – the full-dimensional table.

We demonstrated the implementation of CIPHER for categorical data. The procedure also applies to data with numerical attributes, where the input would be a set of low-dimensional histograms. This implies the numerical attributes will need to be cut into bins first before the application of CIPHER. After the sanitized empirical joint distribution is generated, values of the numerical attributes can be uniformly sampled from the sanitized bins.

For future work, we plan to investigate the theoretical accuracy for CIPHER using some common utility measures (e.g., l_∞ or l_1 errors); to apply CIPHER to data of higher dimensions in terms of both p and the number of levels per attribute to examine the scalability of CIPHER; and to compare CIPHER with

more methods that may also generate differentially private empirical distributions from a set of low-dimensional statistics, such as PrivBayes and the Fourier transform based method, in both data utility and computational costs.

Supplementary Materials

The supplementary materials are posted at https://arxiv.org/abs/1812.05671. The materials contains additional results from experiment 1, more details on the data used in experiment 2 and how $\mathcal{Q}$ is chosen, and the mathematical derivation of the linear equations sets $\boldsymbol{Ax} = \boldsymbol{b}$ for the three-variable and four-variable cases.

Acknowledgments

The authors would like to thank two anonymous reviewers for their comments and suggestions that helped improve the quality of the manuscript.

References

1. Narayanan A, Shmatikov V (2006) How to break anonymity of the netflix prize dataset. CoRR abs/cs/0610105, cs/0610105
2. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: Security and Privacy, 2008. SP 2008. IEEE Symposium on, IEEE, pp 111–125
3. Sweeney L (2013) Matching known patients to health records in washington state data. CoRR abs/1307.1370, 1307.1370
4. Tockar A (2014) Riding with the stars: Passenger privacy in the nyc taxicab dataset. https://research.neustar.biz/author/atockar/
5. Culnane C, Rubinstein BIP, Teague V (2017) Health data in an open world. arXiv preprint arXiv:171205627v1
6. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference, Springer, pp 265–284
7. Dwork C (2008) Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation, Springer, pp 1–19
8. Barak B, Chaudhuri K, Dwork C, Kale S, McSherry F, Talwar K (2007) Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, pp 273–282
9. Chen R, Xiao Q, Zhang Y, Xu J (2015) Differentially private high-dimensional data publication via sampling-based inference. In: Proceedings of the 21th ACM SIGKDD, ACM, pp 129–138
10. Liu F (2016) Model-based differential private data synthesis. arXiv preprint arXiv:160608052

11. Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008) Privacy: Theory meets practice on the map. IEEE ICDE IEEE 24th International Conference pp 277 – 286
12. Bowen CM, Liu F (2020) Comparative study of differentially private data synthesis methods. Statistical Science 35(2):280–307
13. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2014) Privbayes: Private data release via bayesian networks. In: Proceedings of the 2014 ACM SIGMOD, SIGMOD '14, pp 1423–1434
14. Hardt M, Ligett K, McSherry F (2012) A simple and practical algorithm for differentially private data release. In: Advances in Neural Information Processing Systems, pp 2339–2347
15. Liu F, Zhao X, Zhang G (2021) Disclosure risk from homogeneity attack in differentially private frequency distribution. arXiv preprint arXiv:210100311v2
16. McSherry FD (2009) Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, ACM, pp 19–30
17. Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 9(3–4):211–407
18. Liu F (2019) Generalized gaussian mechanism for differential privacy. IEEE Transactions on Knowledge and Data Engineering 31(4):747 – 756
19. McSherry F, Talwar K (2007) Mechanism design via differential privacy. In: Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on, IEEE, pp 94–103
20. Tikhonov AN (1963) On the solution of ill-posed problems and the method of regularization. In: Doklady Akademii Nauk, Russian Academy of Sciences, vol 151 (3), pp 501–504
21. Tikhonov AN, Goncharsky A, Stepanov V, Yagola AG (2013) Numerical methods for the solution of ill-posed problems, vol 328. Springer Science & Business Media
22. Hay M, Rastogi V, Miklau G, Suciu D (2010) Boosting the accuracy of differentially private histograms through consistency. Proceedings of the VLDB Endowment 3(1-2):1021–1032
23. Hay M, Machanavajjhala A, Miklau G, Chen Y, Zhang D (2016) Principled evaluation of differentially private algorithms using dpbench. In: Proceedings of the 2016 International Conference on Management of Data, ACM, pp 139–154
24. Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository. URL http://archive.ics.uci.edu/ml
25. Kim MJ, Han I (2003) The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. Expert Systems with Applications 25(4):637–646