Differentially Private Data Release via Statistical Election to Partition Sequentially

Claire McKay Bowen, Fang Liu, Bingyue Su *

Abstract

Differential Privacy (DP) formalizes privacy in mathematical terms and provides a robust concept for privacy protection. DIfferentially Private Data Synthesis (DIPS) techniques produce and release synthetic individual-level data in the DP framework. One key challenge to developing DIPS methods is preservation of the statistical utility of synthetic data, especially in high-dimensional settings. We propose a new DIPS approach, STatistical Election to Partition Sequentially (STEPS) that partitions data by attributes according to their importance ranks according to either a practical or statistical importance measure. STEPS aims to achieve better original information preservation for the attributes with higher importance ranks and produce thus more useful synthetic data overall. We present an algorithm to implement the STEPS procedure and employ the privacy budget composability to ensure the overall privacy cost is controlled at the pre-specified value. We apply the STEPS procedure to both simulated data and the 2000-2012 Current Population Survey youth voter data. The results suggest STEPS can better preserve the population-level information and the original information for some analyses compared to PrivBayes, a modified Uniform histogram approach, and the flat Laplace sanitizer.

keywords: privacy budget, DIfferentially Private Data Synthesis (DIPS), general utility, propensity score, universal histogram, hierarchical

*Claire McKay Bowen is the Lead Data Scientist for Privacy and Data Security at the Urban Institute. Fang Liu is Professor and Bingyue Su is a doctoral student in the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556. Claire McKay Bowen was supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE-1313583 during part of the development of this paper. Fang Liu was supported by the NSF Grants #1546373, #1717417, and the University of Notre Dame Faculty Research Support Initiation Grant Program. Bingyue Su is supported by the NSF Grant #1717417.

1 Introduction

Researchers have developed various approaches to provide protection for individual sensitive information when releasing data to researchers or the public. Among these approaches, data synthesis is a popular technique that generates synthetic individual-level data given the original data (Drechsler, 2011; Little, 1993; Little et al., 2004; Liu and Little, 2003; Raghunathan et al., 2003; Reiter, 2003, 2009; Rubin, 1993). The data synthesis concept is appealing as it provides surrogate data sets that have the same structure as the original data, and users may perform their own analyses as if they had the original data. A recent data synthesis approach is DIfferentially Private data Synthesis (DIPS) (Bowen and Liu, 2020) that performs data synthesis based on differential privacy (DP), a mathematical framework for controlling privacy risk with a prespecified privacy budget (Dwork et al., 2006).

Some researchers consider DIPS methods as "flat" or one-step in the sense that DP noises are injected all at once or in parallel, such as the smooth and perturbed histograms (Wasserman and Zhou, 2010), MOdel-based DIPS (MODIPS) (Liu, 2016), DPCopula (Li et al., 2014b), and PrivBayes (Zhang et al., 2017), among others. There also exist DP procedures that partition data sequentially and inject random noises in a hierarchical manner to improve accuracy of certain types of queries in low-dimensional settings. Though the focus of these DP mechanisms is not data synthesis, they may be used for data syntheses with some adaptations or with additional steps. We refer to this type of DIPS as sequential or hierarchical as opposed to flat. For example, Xiao et al. (2010) proposed Privelet via a two-step wavelet-based multidimensional partitioning approach to release range count queries. Privelet is effective in the one-dimensional case but makes only slight improvements in the two-dimensional case, and the performance could be even worse at higher dimensions (Qardaji et al., 2013a). Xiao et al. (2012) presented DPCube as a two-phase partitioning approach for data cubes. Gardner et al. (2013) implemented DPCube in biomedical data to explore its practical feasibility on real-world data, but discovered that DPCube is inefficient in constructing accurate high-dimensional histograms. Hay et al. (2010) developed the universal histogram (UH) to inject noise to histogram bi-partitioning with improved accuracy in histogram bin counts by exploring the inherent consistency constraints. Qardaji et al. (2013b) extended the method to relatively high dimensional data. Hay et al. (2016) conducted an extensive comparison on most of the above mentioned algorithms using a set of evaluation principles (DPBench), and provided valuable insights on the pros and cons on the methods for answering 1- and 2-dimensional range queries. Li et al. (2017) introduced the privacy-aware partitioning mechanism and the utility-based partitioning mechanism that depend on a public but personalized privacy parameter. These methods have not been implemented to real-world data likely due to being "unable to provide corresponding error guarantees with such procedures for general functions" (Cummings and Durfee, 2018).

In summary, most of the hierarchical procedures focus on low dimensional data with numerical attributes. To explore the potentials of hierarchical sanitization in improving the utility of synthetic data in relatively high dimensional settings when the original data have both categorical and numerical attributes, we propose a new DIPS procedure – **ST**atistical **E**lection

to **P**artition **S**equentially (STEPS). STEPS injects noises to a hierarchical histogram. The structure of the histogram is informed either by domain or prior knowledge, or by a statistical metric that explores the inherent statistical information in the original data. Different layers on the same branch in the constructed hierarchical histogram contains different attributes. The hierarchical histogram built by STEPS is subject to equality constraints among the nodes from different layers on the same branch; thus a post-processing procedure similar to the UH approach (Hay et al., 2010) is used to ensure that the equality constraints are satisfied. The final step of STEPS is to generate synthetic data from the differentially private hierarchical histogram. We expect that STEPS will improve the statistical utility of the released data compared to data synthesized through a hierarchical histogram with random partitioning given that the data partitioning in STEPS is an informed decision, leveraging knowledge in the data or publicly available relevant knowledge. We apply STEPS to simulated data and the youth voter data from the 2000-2012 Current Population Survey (CPS) and benchmark its utility against some DIPS approaches. To assess the overall utility of the synthetic data, we also develop a propensity score based method that provides a general utility metric and a holistic measure for the similarity between two data sets of the same structure.

The remainder of the paper is organized as follows. Section 2 provides some preliminaries regarding DP and introduces the STEPS procedure. Section 3 applies STEPS to simulated data and compares its capability in preserving the population-level information against PrivBayes and a modified UH procedure. Section 4 implements the STEPS method to the CPS youth voter data and compares the statistical utility of the synthetic data generated by STEPS, PrivBayes, a modified UH procedure, and a flat Laplace sanitizer via several utility analysis. In Section 5, we discuss the implications of our results and provide future research directions.

2 Methodology

2.1 Preliminaries on Differential Privacy

DP provides a mathematical and rigorous framework for protecting individual information in a data set, regardless of the background knowledge or behaviors of data intruders, when releasing queries to the public. Query results, in statistical terminology, are statistics; so we use queries, query results, and statistics, interchangeably in this discuss and denote them by $\mathbf{s}$. We denote the data for privacy protection by $\mathbf{x} = \{x_{ij}\}$ for $i = 1, \ldots, n; j = 1, \ldots, p$. Each row $\mathbf{x}_i$ represents an individual record with p variables/attributes. We assume that the sample size n and the number of attributes p are public knowledge and carry no privacy information.

Definition 1 (**Differential Privacy** (Dwork et al., 2006)). *Let $d(\mathbf{x}, \mathbf{x}') = 1$ represents all possible ways that data set $\mathbf{x}'$ differing from $\mathbf{x}$ by one individual. A sanitization algorithm $\mathcal{R}$ gives ϵ-DP, if for all data sets $(\mathbf{x}, \mathbf{x}')$ that is $d(\mathbf{x}, \mathbf{x}') = 1$ and all result sets $Q \subseteq \mathcal{T}$, where $\mathcal{T}$ denotes the output range of $\mathcal{R}$, to queries/statistics $\mathbf{s}$ that*

$$\left| \log \left(\frac{\Pr(\mathcal{R}(\mathbf{s}, \mathbf{x}) \in Q)}{\Pr(\mathcal{R}(\mathbf{s}', \mathbf{x}') \in Q)} \right) \right| \leq \epsilon, \tag{1}$$

where $\epsilon > 0$ is the privacy loss parameter.

The smaller ϵ is, the more noise are injected to the statistic $\mathbf{s}$ via the sanitized algorithm $\mathcal{R}$; and each individual in the data set has a lower risk of being identified or having their sensitive information disclosed, because $\mathbf{s}$ would be about the same with and without that individual in the data. In addition to the ϵ-DP in Definition 1, there are also conceptual relaxations of the "pure" DP such as the approximate DP (Dwork and Roth, 2013), probabilistic DP (Machanavajjhala et al., 2008), and concentrated DP (Dwork and Rothblum, 2016), among others, so to lessen the amount of noises injected by DP methods by sacrificing a certain amount of privacy.

In regard to what value of ϵ is considered appropriate or acceptable for practical use, Dwork (2008) states the choice of ϵ is a social question. Abowd and Schmutte (2015) acknowledge this and suggest ϵ at 0.01 to ln(3), or even up to 3 in releasing certain statistics in social and economic studies. A wide range of ϵ values have been examined in the literature. For instance, Machanavajjhala et al. (2008) applied DP in the OnTheMap data (commuting patterns of the United States population) and used $(\epsilon = 8.6, \delta = 10^{-5})$-probabilistic DP (a relaxation of the pure DP) to synthesize commuter data. Ding et al. (2011) and Li et al. (2014b) used $\epsilon = 1$ in their experiments. These examples suggest there are many factors that affect the choice of ϵ, including the type of information released to the public, social perception of privacy protection, statistical accuracy of the release data, among others. For a socially acceptable ϵ given a certain type of information, a differentially private mechanism should aim for maximizing the accuracy of released information. In other words, choosing an "appropriate" ϵ is essentially a question of balancing the privacy loss and the accuracy of the released information.

An important property of DP is that the privacy cost increases for every new query released from the same data set, because more information is "leaked" with releasing more query results. Therefore, the data curator must track all statistics calculated on the data set to guarantee the privacy budget does not exceed the prespecified level. For example, if all q queries are sent to data set $\mathbf{x}$, then ϵ/q can be allocated to each query to ensure the privacy budget is maintained at ϵ overall per the *sequential composition* principle (McSherry, 2009). When no overlapping information is requested by different queries, such as when they are calculated from disjoint subsets of a data set, the privacy cost does not accumulate. In such a case, the *parallel composition* principle applies and the overall privacy cost is the maximum privacy budget spent across all the queries (McSherry, 2009).

A common and easy way to implement DP is the Laplace mechanism. A key concept for the Laplace mechanism is the l_1 global sensitivity of statistic $\mathbf{s}$ (either a scalar or a vector) is $\Delta_1 = \max_{\mathbf{x},\mathbf{x}',d(\mathbf{x},\mathbf{x}')=1} \|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}')\|_1$ for all $d(\mathbf{x}, \mathbf{x}') = 1$ (Dwork et al., 2006). Global sensitivity can also be defined in other forms, such as the l_2 global sensitivity (Dwork and Roth, 2013) and l_p global sensitivity for any $p \geq 1$ (Liu, 2019).

Definition 2 (**Laplace mechanism** (Dwork et al., 2006))**.** *The Laplace sanitizer adds noise to statistics $\mathbf{s} = (s_1, \ldots, s_K)$ via $s_k^* = s_k + e_k$, where $e_k \sim Lap(0, \Delta_1/\epsilon)$ and is independent for $k = 1, \ldots, K$ and Δ_1 is the l_1 global sensitivity of $\mathbf{s}$.*

When ϵ is small or Δ_1 is large, more Laplace noise is added to $\mathbf{s}$. Other common DP mechanisms include the Gaussian mechanism that is built upon the approximate or the probabilistic DP (Dwork and Roth, 2013; Liu, 2019), and the exponential mechanism that can release both numerical and non-numerical queries (McSherry and Talwar, 2007), among others. The definition of the exponential mechanism is presented below, which will be used in our proposed STEPS procedure.

Definition 3 (**exponential mechanism** (McSherry and Talwar, 2007))**.** *Let u be a utility function that assigns a score to each possible output of a query. The exponential mechanism that satisfies ϵ-DP releases query result s calculated from data D with probability* $\exp(u(s)\frac{\epsilon}{2\delta_u}) \Big/ \int_{s'} \exp\Big(u(s')\frac{\epsilon}{2\delta_u}\Big)\, ds'$*, where δ_u is the maximum change in score u with one element change in data D.*

2.2 STatistical Election to Partition Sequentially (STEPS)

2.2.1 Overview

STEPS aims at privately selecting a decomposition sequence of the joint distribution of p attributes $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ in the data, and sanitizing each component in the decomposition sequentially to achieve better original information preservation for the attributes that are more important to the practical or statistical understanding of the population-level signals in the data. The outcome of STEPS is a private joint distribution $f(\mathbf{X})$, from which synthetic data can be generated and released.

The decomposition of the joint distribution can be a "full" decomposition such as $f(\mathbf{X}) = f(X_1)f(X_2|X_1)\ldots, f(X_p|X_1, \ldots, X_{p-1})$, or "partial" decomposition such as $f(\mathbf{X}) = f(X_1)$ $f(X_2|X_1)f(X_3, \ldots, X_{p-1}|X_1, X_2)$ or $f(\mathbf{X}) = f(X_1, \ldots, X_j|X_{j+1}, \ldots, X_p)f(X_{j+1}, \ldots, X_p)$, etc. For example, suppose a data set has three attributes X_1, X_2, X_3, and the decomposition sequence chosen by STEPS is $f(X_3)f(X_1|X_3)f(X_2|X_1, X_3)$. STEPS sanitizes the three components $f(X_3)$, $f(X_1|X_3)$, and $f(X_2|X_1, X_3)$, independently, taking into account the necessary equality constraints along the way. Since all three components contain information on X_3, STEPS will aggregate all the sanitized information on X_3 when generating the private joint distribution; similarly for X_1, the information on which is contained in two queries ($f(X_1|X_3)$ and $f(X_2|X_1, X_3)$), and STEPS will aggregate both sources of sanitized information on X_1 when generating the private joint distribution. Suppose that the sanitization of each component is allocated the same amount of privacy budget, since X_3 is involved in all three components in the decomposition, it receives the most amount of privacy budget collectively and its original information is expected to be better preserved than the other two; so does X_1 compared to X_2.

2.2.2 Procedure

The STEPS procedure has three main steps: data partitioning and hierarchical tree construction; sanitization and correction; and synthesis and release. While STEPS can deal with both categorical and numerical attributes, the numerical ones will be first cut into his-

togram bins and be treated as categorical attributes afterwards until the last step of synthetic data generation.

In the first step of data partitioning and hierarchical tree construction, STEPS selects a decomposition of the joint distribution $f(\mathbf{X})$. The decomposition sequence, once decided, can be represented in the format of a hierarchical tree $\mathcal{T}$. The closer an attribute is placed to the top of the tree, more statistics that involve that attribute will be sanitized the subsequent steps, the more privacy budget the attribute will receive collectively, and thus more original information is expected to be preserved for that attribute. Figure 1 shows an example of $\mathcal{T}$ built by STEPS when $p = 4$. The layers are denoted by $l = 0, 1, \ldots, L = 4$, and the nodes in layer l are denoted by $\mathbf{v}^{(l)}$ and each of them contains the subset of data following the partitioning rule till that node along its branch of the tree. $\mathbf{v}^{(0)}$, the node at the top of $\mathcal{T}$ contains the whole data set. K_j represents the number of levels of variable X_j for $j = 1, \ldots, 4$. Note that the order of the variables for partitioning may differ by branch, and the branches do not have to be of the same length. The decomposition in the tree in Figure 1 is $f(\mathbf{X}) = f(X_3, X_4|X_2 = 1, X_1 = 1)f(X_2 = 1|X_1 = 1)f(X_1 = 1) + \ldots + \sum_{j=1}^{K_4} f(X_3|X_2 = K_2, X_1 = 1, X_4 = j)f(X_4 = j|X_2 = K_2, X_1 = 1)f(X_2 = K_2|X_1 = 1)f(X_1 = 1) + \ldots + f(X_2, X_3|X_1 = K_1, X_4 = 1)f(X_4 = 1|X_1 = K_1)f(X_1 = K_1) + \sum_{j=1}^{K_2} f(X_3|X_4 = K_4, X_1 = 1, X_2 = j)f(X_2 = j|X_4 = K_4, X_1 = K_1)f(X_4 = K_4|X_1 = K_1)f(X_1 = K_1)$.

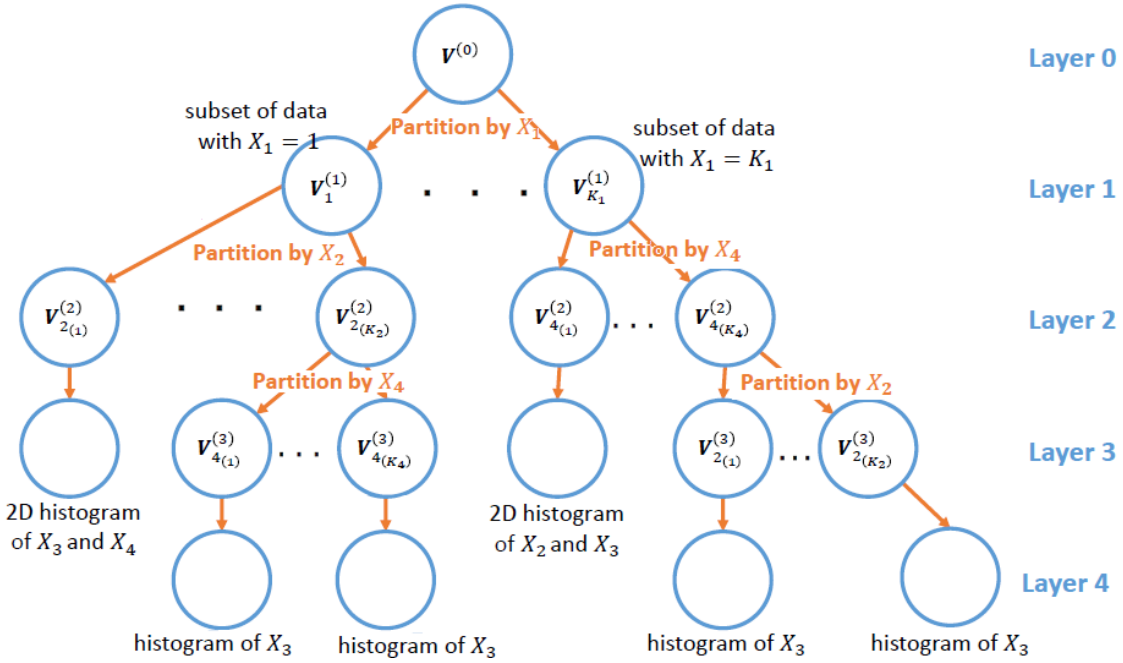


Figure 1: Illustration of an hierarchical tree built by STEPS in a 4-variable case.

The sequence by which the data are partitioned or by which the attributes are aligned and placed on trees can be determined in at least two ways. First, one can leverage prior information or domain knowledge regarding the importance of the attributes, either practical or statistical, in a particular data set. The attributes that are of more interest to practitioners or more important per the prior knowledge will be placed closer to the root (top) of the tree.

Applying external knowledge to aid certain aspects of developing a DP procedure (e.g., hyper-parameter tuning) has been used in other privacy-preserving approaches to help save on privacy cost and improve the utility of released results. This approach does not cost any privacy of the current data as the information comes from outside. For the second approach, one may apply a metric that measures the importance of each attribute to the statistical understanding the original data, such as a quantity that quantifies the contribution of an attribute in explaining the variability of the data; attributes that explain more variability will be placed closer to the root of the tree. For example, information-theoretic model selection criteria such as AIC and BIC can be used to determine the then the attribute selected to partition the data in a parent node is the one with the smallest AIC among all the univariate log-linear models fitted to the data. Since the metric is calculate from the original data and determines the tree structure, a portion of the total privacy budget should be allocated to this step to keep the determination of the partitioning sequence private.

In the second step of sanitization and correction, a differentially private mechanism injects noises to the count of the data points contained in each of the nodes of the constructed tree $\mathcal{T}$. After the independent noise injection, the sum of the counts in the children nodes (denoted by $\text{succ}(v)$) are very likely not equal to the count in their parent node v, but they should. For example, a node v contains all the data points whose attribute "race" is "white". Say its original count of 100, and its children $\text{succ}(v)$ from further partitioning by "gender" include "white female" with an original count of 44 and "white male" of with an original count of 56. Suppose the sanitized counts are 120, 48, and 50, respectively, then "white male" and "white female" no longer add up to "white". To ensure the equality constraints between $\sum \text{succ}(v)$ and v in the hierarchical decomposition, a post-processing procedure similar to the UH approach is applied, with some necessary modifications. Specifically, z is first calculated via Eqn (2) as a weighted average of the directly sanitized count $n^*[v]$ and the indirectly sanitized count $\sum_{u \in \text{succ}(v)} z[u]$ summed from its children nodes in a bottom-up manner, and then the inconsistency is corrected to obtain the final consistent counts $\bar{n}^*$ via Eqn (3).

$$z[v^{(l)}] = \begin{cases} n^*[v^{(l)}], & \text{if } l = L \text{ (v is a leaf node)} \\ \frac{b^{L-l} - b^{L-l-1}}{b^{L-l}-1} n^*[v^{(l)}] + \frac{b^{L-l-1}-1}{b^{L-l}-1} \sum_{u \in \text{succ}(v^{(l)})} z[u] & \text{if } l = L-1, \ldots, 1 \end{cases}; \quad (2)$$

$$\bar{n}^*[v] = \begin{cases} z[v^{(l)}], & \text{if } l = L \text{ (v is a leaf node)} \\ z[v^{(l)}] + \frac{1}{b}(\bar{n}^*[w] - \sum_{u \in \text{succ}(w)} z[u]) & \text{if } l = L-1, \ldots, 1 \end{cases}. \quad (3)$$

w in Eqn (3) represents v's parent; b in Eqns (2) and (3) is the number of children per node, assumed to be the same across all nodes in $\mathcal{T}$ in the UH approach (Hay et al., 2010), and can be set by the data user ($b \geq 2$); b^l denotes b to the l power in Eqn (2). A tree built by STEPS might have branches of different length (e.g. Figure 1). To accommodate the requirement of the same b across all nodes so that Eqns (2) and (3) still apply, *phantom* categories can be generated to make each attribute have the same number of categories as

the attribute with the most categories. The "original" counts in the phantom categories are 0, and will remain at 0 during the sanitization and correction process.

In the last step of synthesis and release, if all attributes are categorical, we can release directly the sanitized counts of the nodes in the bottom layer of the tree; otherwise, we can apply uniform sampling to draw synthetic values for the numerical attributes in each bin of the formed histograms in each node.

2.2.3 Algorithm

The algorithmic steps of the STEPS procedure are given in Algorithm 1.

Algorithm 1 STatistical Election to Partition Sequentially with L layers (STEPS-L)

Input: number of layers L ($\leq p$) (layer 0 is not counted toward L which contains all data points); overall privacy budget ϵ; the portion $r\epsilon$ for $r \in [0, 1)$ allocated for determining decomposition; utility function u (e.g., AIC) for the exponential mechanism and its sensitivity δ_u; number of synthetic sets m; original data D.
Output: m synthetic sets $\tilde{D}^{(1)}, \ldots, \tilde{D}^{(m)}$
If there is a predefined partitioning sequence, $r = 0$ and build tree $\mathcal{T}$ according to the predefined sequence.
Else define $\mathcal{A}^{(0)}_{v^{(0)}[1]}$ that includes all p attributes in D.
  For $l = 0$ to L
    For $k = 1$ to $K^{(l)}$ ($K^{(l)}$ is number of nodes in layer l)
    • Apply a univariate loglinear model to the data contained in node $\mathbf{v}^{(l)}[k]$ on each attribute from its availability set $\mathcal{A}^{(l)}_{\mathbf{v}^{(l)}[k]}$, and calculate u.
    • Choose a partitioning variable via the exponential mechanism with privacy budget $r\epsilon/(mL)$. Denote the selected attribute by $X_j[k]$.
    • Exclude $X_j[k]$ from the availability sets for the all children nodes of $\mathbf{v}^{(l)}[k]$ in layer $l + 1$.
    End For
  Obtain the full histogram over all the attributes in $\mathcal{A}^{(L)}_{\mathbf{v}^{(L)}}$ for all nodes in layer L.
  End For
End Else
Sanitize all nodes in the generated tree $\mathcal{T}$ (e.g. via the Laplace mechanism), and apply the inconsistency correction in Eqns (2) and (3) to obtain the final sanitized counts $\bar{\mathbf{n}}^*$.
If all attributes are categorical, release the sanitized counts of the nodes in the bottom layer of $\mathcal{T}$; otherwise, apply uniform sampling to draw synthetic values for the numerical attributes on which histograms are formed.

The input to the STEPS algorithm includes a user-specified number for the partition layers L. Though the case of $L > p$ would be possible by allowing the levels of an attribute to span across multiple layers, this does not seem necessary from a data synthesis perspective, especially when the partitioning sequence is determined statistically. Therefore, we require that a variable used in partitioning data in earlier layers is no longer available along that

branch; in other words, the set of available attributes $\mathcal{A}^{(l)}_{\mathbf{v}^{(l)}[k]}$ for partitioning the k-th node $\mathbf{v}^{(l)}[k]$ in layer l will shrink as the tree grows. When $L < p$, there are more than one attributes in $\mathcal{A}^{(L)}_{v^{(L)}[k]}$ for the k-th nodes in layer L, but no more partitioning will be applied and the leaf nodes will be made of the cells from the full cross-tabulation over all the remaining attributes in $\mathcal{A}^{(L)}_{\mathbf{v}^{(L)}[k]}$.

Regarding the choice of L, a larger L is associated with more computational complexity and less privacy budget per layer. On the other hand, since the counts in then top layers are weighted averages per Eqn. (3), the decrease in the privacy budget per layer could be offset or even trumped by the information gain aggregated over multiple sources of information. Given the above reasoning, we conjecture that there exists an optimal L based on the trade-off between privacy budget and data utility. Users may try different values of L and pick a L leads to the best utility among all available choices; but the procedure of choosing L itself, if using the original data, costs privacy.

If the tree structure is suggested by external knowledge, then the tree building costs no privacy; if the data set itself is used to suggest the structure of a tree, the overall privacy budget will be split between building and sanitizing the hierarchical tree. Specifically, a certain portion r of the overall privacy, which is further split into L layers, will be allocated to find the optimal split variable for a node in each layer via the Exponential mechanism; the rest of the privacy budget $(1-r)$ can be used to sanitize the node counts in the constructed tree, which is also further split into L partition layers. The optimal r, a hyperparameter, likely depends on the data; but users could preset a r value they are willing to spend on tree building if they do not want to spend budget to pick r. Since different nodes in the same layer do not have overlapping information, choosing the split variables and sanitation of the nodes follow the parallel composition principle. Taken together, with m synthetic data sets, and if each layer receives the same amount of budget for count sanitization, then each node count in layer l receives a budget of $(1-r)\epsilon/(mL)$.

When the exponential mechanism is used to privately choose a partitioning variable from the availability set $\mathcal{A}^{(l)}_{\mathbf{v}^{(l)}[k]}$ for node $\mathbf{v}^{(l)}[k]$, it samples attribute $j \in \mathcal{A}^{(l)}_{\mathbf{v}^{(l)}[k]}$ with probability

$$\frac{\exp\left(u_j(\mathbf{v}^{(l)}[k])\epsilon^{(l)}_k/(2\delta_u)\right)}{\sum_{j' \in \mathcal{A}^{(l)}_{\mathbf{v}^{(l)}[k]}} \exp\left(u_{j'}(\mathbf{v}^{(l)}[k])\epsilon/(2\delta_u)\right)}, \tag{4}$$

where δ_u is the maximum change in the utility function u with one element change in the data contained in $\mathbf{v}^{(l)}[k]$, and $\epsilon^{(l)}_k$ is the privacy budget node $\mathbf{v}^{(l)}[k]$ receives for employing the exponential mechanism. Users may choose any reasonable utility function u in Algorithm 1, but an easy and useful choice is AIC, whose $\delta_u = 2$ per Lemma 1.

Lemma 1. The sensitivity δ_u is 2 when the utility function u in the exponential mechanism is AIC from a univariate log-linear model with a single attribute.

Proof. $\text{AIC} = -2\log(\mathcal{L}) + 2K$, where $\mathcal{L}$ is the likelihood function and K is the number of the parameters of the model. For the univariate log-linear model with an independent variable

of K levels,

$$\mathcal{L} = \frac{n_v!}{\prod_{k=1}^{K} n_{v,k}!} \prod_{k=1}^{K} p_k^{n_{v,k}}; \text{ and } \log(\mathcal{L}) = \log(n_v!) - \sum_{k=1}^{K} \log(n_{v,k}!) + \sum_{k=1}^{K} n_{v,k} \log(p_k), \quad (5)$$

where n_v is the data size, and $n_{k,v}$ is the count in level k of that attribute. Removing one element from v leads to a loss of one observation in one of the K categories, say level j. The likelihood function based on the $n_v - 1$ observation is

$$\mathcal{L}' = \frac{(n_v-1)!}{(n_{v,j}-1)! \prod_{k\neq j} n_{v,k}!} p_j^{n_{v,j}-1} \prod_{k\neq j} p_k^{n_{v,k}}; \text{ and}$$

$$\log(\mathcal{L}') = \log((n_v-1)!) - \log((n_{v,j}-1)!) - \sum_{k\neq j} \log(n_{v,k}!) + n_k \log(p_k) + \sum_{k=1}^{K} n_{v,k} \log(p_k). \quad (6)$$

The change in AIC is $\Delta\text{AIC} = -2\log(\mathcal{L}) + 2K + 2\log(\mathcal{L}') - 2K'$ with the removal of one attribute. Plugging the log-likelihoods from Eqns (5) and (6) in the right side of ΔAIC, we have $\Delta\text{AIC} = \log(n_v) - \log(n_{v,j}) + n_{n,j}\log(p_j) - (n_{n,j}-1)\log(p_j) + 2K - 2K'$. The true parameter p_j is unknown and its maximum likelihood estimate is $n_{v,j}/n_v$ and $(n_{v,j}-1)/(n_v-1)$ before and after removing an observation. Therefore, $\Delta\text{AIC} = \log(n_v) - \log(n_{v,j}) + n_{v,j}\log(n_{v,j}/n_v) - (n_{n,j}-1)\log((n_{v,j}-1)/(n_v-1)) + 2K - 2K' = (n_{v,j}-1)\log(n_{v,j}(n_v-1)/(n_v(n_{v,j}-1))) + 2K - 2K'$. When $n_j \geq 2$, $K = K'$ and thus $\Delta\text{AIC} = (n_{v,j}-1)\log(n_{v,j}(n_v-1)/(n_v(n_{v,j}-1))) \in (0,1)$. When $n_j = 1$, then $K' = K-1$, and $\Delta\text{AIC} = 2$. Taken together, δ_u with AIC is 2. ■

Setting $\delta_u = 2$ (from Lemma 1) in Eqn (4), then attribute j is sampled with probability

$$\exp(-\text{AIC}_j \epsilon_k^{(l)}/4) / \textstyle\sum_{j'} \exp\left(-\text{AIC}_{j'} \epsilon_k^{(l)}/4\right) \quad (7)$$

If an AIC has a large negative value, direct application of Eqn (7) may lead to overflow problems in computation. It is thus suggested to replace AIC_j and $\text{AIC}_{j'}$ above by their differences from $\max_{j'} \text{AIC}_{j'}$ in practical implementation.

Algorithm 1 may have other variants. One such variant is that the data in a node can be partitioned by a group of variables rather than by a single variable, if that group of variables are similar in their importance measures. For example, suppose there are 5 variables in a data set, each associated with their respective importance measure. Suppose L is pre-set at 2. There are 15 ways of splitting the 5 variables into two clusters if each cluster has to have at least one variable. To choose which splitting scheme to use, we may use the l_1 difference between the two clusters on the average importance measures across the variables in each cluster as the utility function for the exponential mechanism. The clustering scheme with the largest l_1 difference has a higher probability being chosen. Once the exponential mechanism chooses a clustering, the group with the higher average importance measure will

be put in layer 1 and the other will be in layer 2. This generalized splitting scheme is what is used in the simulation studies in Sec 3.

Algorithm 1 suggests releasing multiple synthetic sets ($m > 1$) so to capture the sanitization and synthesis uncertainty to yield valid statistical inferences based on the released synthetic data. While releasing multiple sets is not the only way to capture the uncertainty, it is likely the most straightforward and simplistic way in terms of practical implementation. Statistical inference for analysis over the multiple sets can be obtained via the formulas given in Liu (2016) and Bowen and Liu (2020). To preserve the overall DP in the release of m sets, each synthetic set is allocated $1/m$ of the total privacy budget ϵ per the sequential composition. In terms of the choice for m, m too large will lead too much information loss for a single synthetic set to be remedied by aggregating over m sets of information; and m too small might not adequately propagate the inherent uncertainty from sanitization and synthesis. Our empirical studies suggest $m = 3$ to 6 is a good choice (Liu, 2016).

2.3 STEPS vs UH and PrivBayes

The STEPS procedure focuses on constructing a differentially private joint distribution, in relatively high dimensional settings, from which synthetic data are generated. It aims at preserving more information for important variables, where "important" can be defined either statistically or per domain knowledge. By contrast, most existing partitioning approaches focus on improving the accuracy of marginal counts in low-dimensional setting. In addition, they often allow the same attribute to span multiple layers whereas different layers in STEPS contain non-overlapping attributes.

Among the existing partitioning-based DP approaches, STEPS relates to the UH the most as it uses the inconsistency correction rules developed in the latter. Exploring and maintaining the inherent consistency constraints in the UH procedure helps improve the accuracy of low-dimensional marginal queries. The UH is mostly studied in the context of a single attribute, where the high-level nodes present large-range queries and the leaf nodes in the lowest level are the finest categories/bins for the attribute. Though it can be applied to multidimensional data, its benefit over the one-step Laplace sanitizer seems to diminish over increasing dimensionality (Qardaji et al., 2013a,b). In addition, the UH approach has an exponential time complexity $O(b^L)$ in L for a given b (Eqns (2) and (3)), implying a large L would dramatically increase the computational time (Qardaji et al., 2013b).

Another related DIPS method to STEPS is PrivBayes (Zhang et al., 2017), which also relies on the decomposition of the full distribution $f(\mathbf{X})$ of the attributes in the data. However, PrivBayes is different from STEPS in several aspects. First, PrivBayes uses a low-dimensional distribution to approximate the full distribution $f(\mathbf{X})$ by exploring the conditional independence among the attributes. In other words, PrivBayes is model-based approach (though Bayesian networks) to represent the signals in the data. As such, the sanitized data are subject to bias if the Bayesian network does not provide a good fit to the data. In contrast, STEPS does not imposes a model on the data and it always sanitizes the full-dimensional empirical distribution, though it may leverage modelling to determine which

layer to put a certain attribute and emphasizes the preservation of more information on more important attributes. Second, the PrivBayes injects noises to all parent-children node pairs in parallel to derive the differentially private approximate probability mass function to $f(\mathbf{X})$ from which synthetic data can be generated; correction for inconsistency is not needed after the sanitation.

2.4 A General Utility Metric

To assess the similarity between synthetic data and actual data, we develop the SPECKS (**S**ynthetic data generation; **P**ropensity score matching; **E**mpirical **C**omparison via the **K**olmogorov-**S**mirnov distance) metric. SPECKS is a propensity-score-based general utility measure and can be used to compare the similarity of two data sets of the same structure of any dimension without making assumptions on the distributions of the attributes. The procedure of SPECKS is given below. Each step is straightforward and easy to implement.

1) Combine the original and synthetic data, each of size n. Create an indicator variable T where $T_i = 1$ if record i is from the synthetic data and $T_i = 0$ otherwise for $i = 1, \ldots, 2n$.

2) Calculate the propensity score for each record i, $e_i = \Pr(T_i = 1|\mathbf{x}_i)$, through a classification algorithm, with the data attributes as input features.

3) Calculate the empirical CDFs of the propensity score, $\hat{F}(e)$ and $\tilde{F}(e)$, for the actual and the synthetic groups, separately.

4) Compute the Kolmogorov-Smirnov (KS) distance $d = \sup_e |\tilde{F}(e) - \hat{F}(e)|$ between the two empirical CDFs (if multiple synthetic data sets are generated, the average KS distance over the multiple sets is taken).

If the synthetic data preserve the original information well, then the observations from the two groups are indistinguishable and a small KS distance between the original and synthetic empirical CDFs is expected. In the second step of the SPECKS procedure, any classifier (e.g., logistic regression, random forests, SVM) can be used. In the case of logistic regression, the model covariates may include the main effects of the data attributes or interaction terms among the attributes. This implies that the propensity scores will vary by classifier. Bowen and Snoke (2019) note that different classifiers measure different types of distributional similarity. For instance, a logistic regression model with only main effects measures data similarity in the marginal distributions (simultaneously) whereas a more complex CART model can measure similarity in high-order dependency among the attributes.

Compared to other propensity-score-based utility measures or discriminate-based methods, SPECKS has similar steps such as the calculation of propensity scores, but differs in how it formulates the final utility metric from the estimated propensity scores (Steps 3 and 4 above). In Sakshaug and Raghunathan (2010), the propensity scores are discretized based on how the Chi-squared test is formulated. Woo et al. (2009) calculate the mean squared error (MSE) of the calculated propensity score vs the true proportion of synthetic cases. Snoke et al. (2018) normalize the MSE statistic by its expected null value and standard deviation, which helps with its interpretability and also seems to be more sensitive in telling

the synthetic data apart from the actual data. However, the derivation of the expected null value and standard deviation is driven large-sample assumptions. In contrast, SPECKS utilizes the KS distance which is the maximum distance of two empirical CDFs, and thus considers the worst-case separation between the synthetic data and the actual data.

3 Simulation Studies

In this section, we implement the STEPS method in simulated data to generate differentially private synthetic data, and compare the statistical utility of the synthesized data between STEPS vs. a modified version of the UH and PrivBayes.

3.1 Choice of Methods for Comparison

The reasons for choosing the UH and PrivBayes to compare with STEPS are given below. The same justifications apply to the case study in Sec 4.

First, the UH and PrivBayes methods are the most related methods to STEPS as presented in Sec 2.3. STEPS uses the inconsistency correction rule of the UH In addition, Qardaji et al. (2013b) suggest that the UH generally is the best data-independent algorithm and achieves lower error for range queries from one-dimensional histograms, compared to many data-independent methods. When implementing the procedure, we made some modification to the original UH to accommodate computational constraints and to handle multi-dimensional data. Both STEPS and PrivBayes build a differentially private joint distribution among the data attributes from which synthetic data are generated. Similar to STEPS, PrivBayes deals with numerical attributes by discretizing them into histogram bins.

Though Privelet and DPcube, originally developed for answering range queries from one- or two- dimensional histograms, can be potentially used for data synthesis, we do not examine them in the simulation studies as neither offers better performance than the UH approach in general even in the low-dimensional setting (Hay et al., 2016). MWEM (Hardt et al., 2012) aims to obtain differentially private queries rather than generating synthetic data from an empirical distribution. To use MWEM for data synthesis, choosing a good set of queries that represent the information of the original data is vital. Even with a good set of queries to start with, it does not necessarily outperform the flat Laplace sanitizer when generating synthetic data (Eugenio and Liu, 2018). In addition, MWEM is proved to be inconsistent and its performance highly depends on the number of iterations (Eugenio and Liu, 2018; Hay et al., 2016; Li et al., 2014a). Though the DAWA method (Li et al., 2014a) beats other methods including DPcube, MWEM, and Privelet for low-dimensional range queries, the good performance replies on allocating some privacy budget to identify an optimal bucketing scheme on the numerical attributes before the sanitization. Given most of the attributes in our simulation and case studies are either categorical or ordinal with limited number of levels, the advantages of DAWA will not be brought into full strength. Though the model-dependent MODIPS method can deal with all types of data in theory, it has a couple of limitations for practical implementation, such as that not every sufficient statistic is easy to sanitize and a mis-specified model would generate biased synthetic data.

DPcopula uses copula to model pairwise dependency among attributes, based on which data are synthesized; the method assumes continuous marginal CDF; and the synthetic data can also be sensitive to the type of copula employed.

3.2 Simulation Setting

We simulated data from two linear regression models. For each model, we simulated 200 data sets. Model 1 is main-effect model: $Y = \beta_0 + \sum_{j=1}^{10} \beta_j X_j + \epsilon$ with $\epsilon \sim N(0, 1)$. X_1 to X_4 are categorical predictors with 2, 3, 4 and 5 levels respectively, and the true $\boldsymbol{\beta}$ values are $0, 0.5, (0.5, -0.8), (1, 0.5, 0.5), (0.5, 0.7, 0.8, 0.5)$, respectively. Model 2 is a saturated model: $Y = \boldsymbol{\beta}^T \mathbf{X}_{1234} + \epsilon$ with $\epsilon \sim N(0, 1)$. $\mathbf{X}_{1234}$ refers the 4-way interaction term among X_1, X_2, X_3, X_4, and $\boldsymbol{\beta}$ is 120-dimensional with each of its elements sampled from a standard normal distribution and then fixed for all 200 repeats. The two models represent the two extremes of a wide spectrum of possible models with 4 predictors.

We examined two sample sizes scenarios, $n = 10,000$ and $n = 4,000$, in each model. We set privacy budget ϵ at $0.5, 2, 5, 10$, respectively. For each of the 12 simulation scenarios (two models at two sample sizes with three DIPS methods), four sets of synthetic data were generated by each of the three DIPS approaches (STEPS, UH, and PrivBayes). The numerical attribute Y was cut into 15 bins before the application of each method, and was treated as categorical until the last step before data release (refer to line 18 of Algorithm 1).

For STEPS, we split the total ϵ in a $1:9$ ratio between building the tree and sanitizing node counts for the tree. We used AIC as the utility function in the exponential mechanism for choosing partitioning attributes. Since the group of variables (X_1, X_2, X_4) had much smaller AIC than the group (X_3, Y), we used the former group as the partitioning variables in Layer 1, and put the second group (X_3, Y) in Layer 2 to form a tree with $L = 2$. The Laplace mechanism was applied to sanitize the node counts in the tree. In addition, We created phantom categories for X_1 to X_4 to match the number of bins b in discretized Y so that the inconsistency correction formulas in Eqns (2) and (3) can be applied (see Sec 2.2.2). For the UH, we modified the original approach to allow unequal number of children per node. In fact, the modified UH is more of a STEPS procedure with a random partitioning sequence and $L = 4$. For PrivBayes, we applied the Python codes by Ping et al. (2017), available on GitHub (Ping, 2018). The degree of the Bayesian network was set at 2 in all the simulation scenarios. The total ϵ was split in half between choosing a Bayesian network vs. sanitizing the joint distribution among the attributes (the default setting).

We run the true underlying linear regression models on each synthetic set and calculated the inferences on the regression coefficients using the combination rule given in Liu (2016) and Bowen and Liu (2020). We define the parameter of interest as β and the estimate of β in the j^{th} synthetic data by $\hat{\beta}_j$ and the associated standard error by SE_j. The final point estimate $\bar{\beta}$ is

$$\bar{\beta} = m^{-1} \sum_{d=1}^{m} \hat{\beta}_j \tag{8}$$

with $\text{Var}(\bar{\beta})$ estimated by

$$T = m^{-1} B + W, \tag{9}$$

where $B = \sum_{j=1}^{m}(\hat{\beta}_j - \bar{\beta})^2/(m-1)$ (between-set variability) and $W = m^{-1}\sum_{j=1}^{m} SE_j^2$ (average per-set variability); and tests and confidence intervals are based on

$$(\bar{\beta} - \beta)T^{-1/2} \sim t_{\nu=(m-1)(1+mW/B)^2}. \tag{10}$$

For Model 2, the ridge regression was applied given the large number of parameters. The bias, root mean squared error (RMSE), and coverage probability (CP) of the 95% confidence interval (CI) were obtained for each regression coefficient in each model. We also applied the estimated private models based on the synthetic data to predict Y in an independent testing data set of $n = 100$, and reported the prediction MSE.

3.3 Results

The results are presented in Figures 2 and 3 for the main-effect model (Model 1) and the saturated model (Model 2), respectively.

When the true model contains only the main effects, PrivBayes performs the best in parameter estimation bias, CP, and prediction RMSE and the worst in parameter estimation RMSE; STEPS performs the best in parameter estimation RMSE; the modified UH is the worst in parameter estimation bias, CP, and the prediction RMSE. As expected, the inferences improve as n or ϵ increases; the only exception is the trend of CP across ϵ for PrivBayes at both n scenarios, where more deviation from the nominal 95% level is observed as ϵ gets larger. This counter-intuitive observation is likely due to the Y discretization. Specifically, when ϵ is large, the main source of information loss comes from the (deterministic) discretization in Y and the main source of variation across the multiple synthetic sets is the uniform sampling from each sanitized bin, rather than the noises introduced by sanitization. When ϵ is small, the main source of information loss and variability across multiple synthetic sets is sanitization, and releasing multiple synthetic sets is designed to take into account that source of variation and the CP is rather better than that at large ϵ.

When the true model is saturated, STEPS performs the best overall; PrivBayes is the worst in parameter estimation RMSE, and the modified UH is the worst in CP and prediction RMSE. For STEPS, the inferences improve as n or ϵ increases as expected, with the only exception in the case of CP across all ϵ values at both n scenarios – likely due to the same reason as discussed above in the case of PrivBayes in the main-effect model. For the modified UH, the inferences get better as n increases, but the trend is not obvious across all ϵ values. For PrivBayes, the trends across n and ϵ are not obvious in any of the examined metrics.

4 Application to the 2002-2012 CPS Youth Voter Data

4.1 Data Description

The 2000-2012 Current Population Survey (CPS) is downloaded from the Harvard Dataverse. The CPS is the primary source of labor force statistics for the United States population. In election years, the CPS also collects data on reported voting and registration,

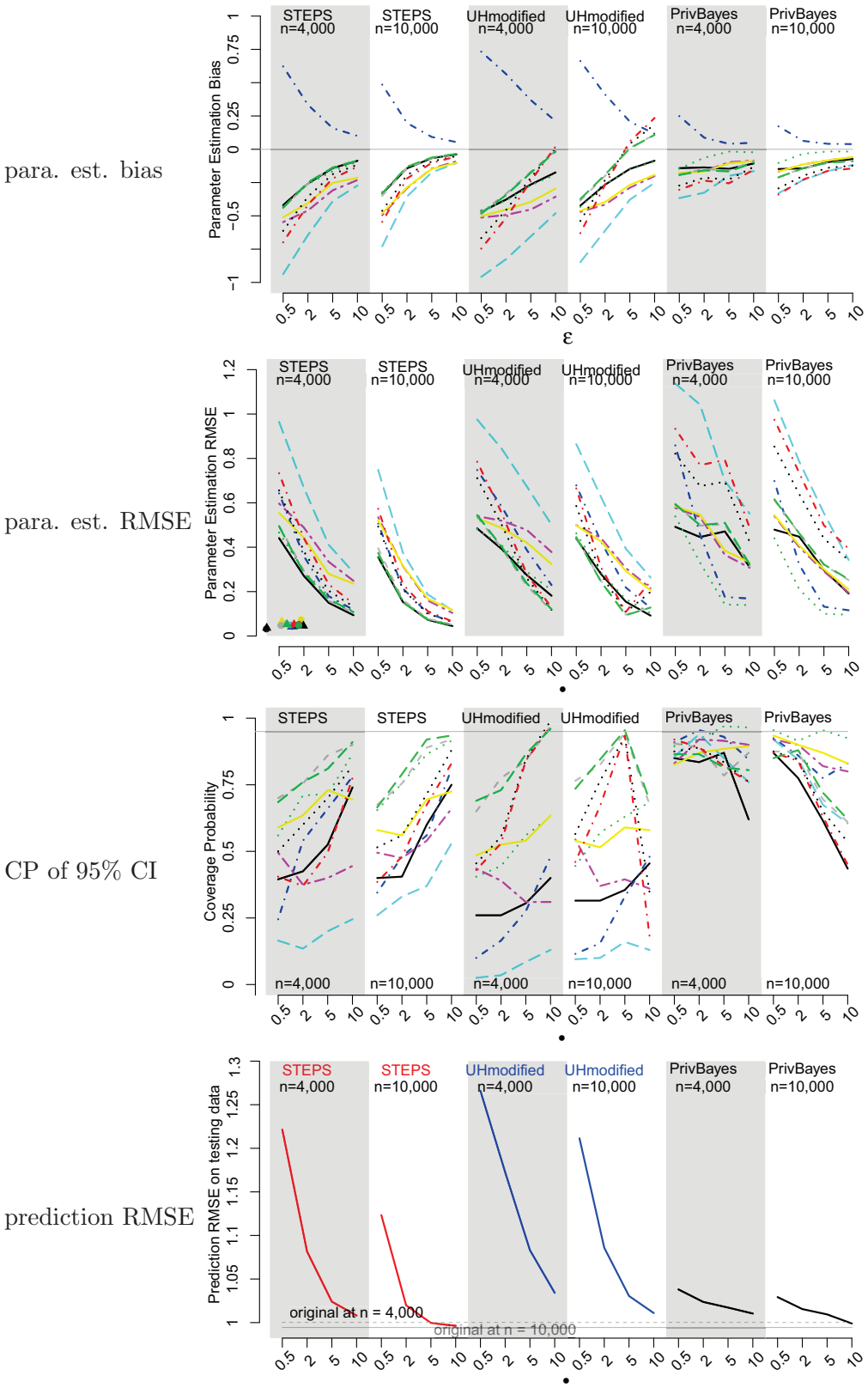
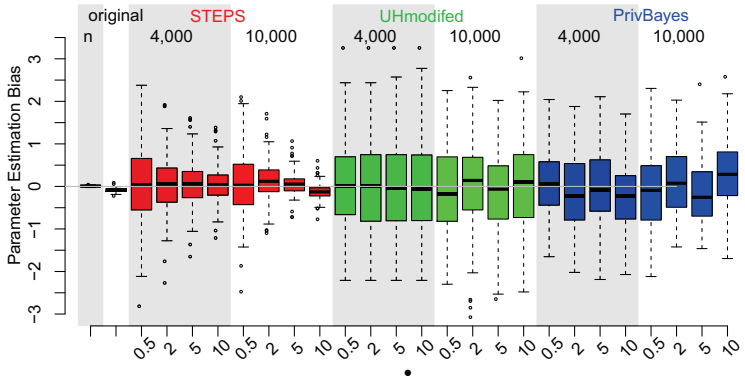
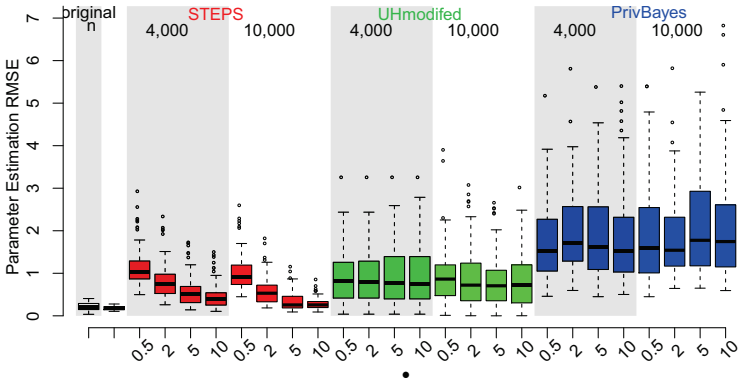


Figure 2: Inferences based on differentially private synthetic data in the simulation study for Model 1. Each line represents a different regression coefficient in Model 1 in each of the top 3 plots.

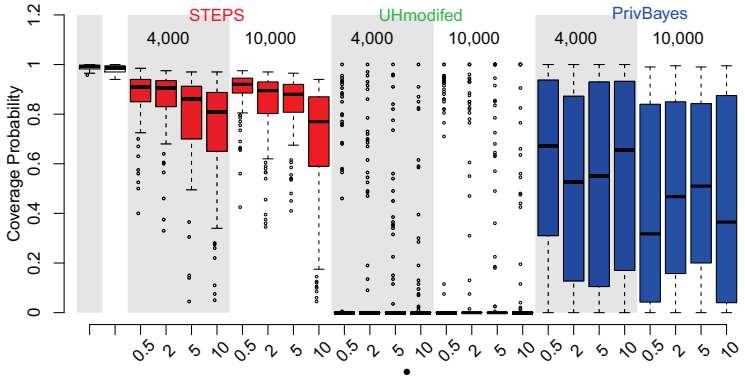
para. est. bias



para. est. RMSE



CP of 95% CI



prediction RMSE

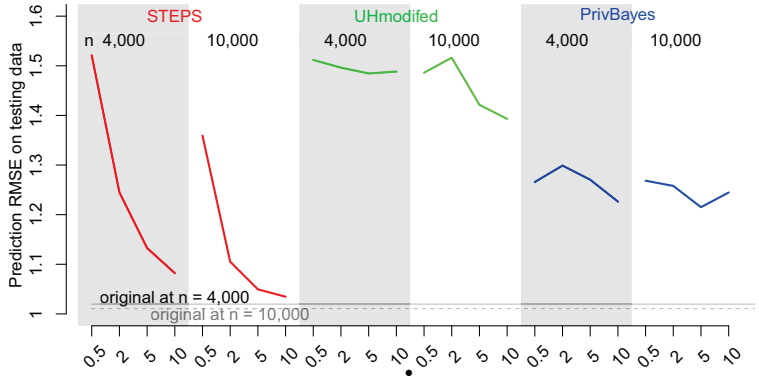


Figure 3: Inferences based on differentially private synthetic data in the simulation study for Model 2. In the bias figure, each box plot represents the distribution of the biases over the 120 model parameters in a specific simulation scenario; similarly for the RMSE and CI.

providing a large and nationally representative sample with coverage of both registered and non-registered individuals, and reports statistics on the voter turnout, age, race, among others. The voting and registration data from the CPS are frequently used and cited by various news outlets such as Time (ABRAMS, 2019), New York Times (Hill and Grumbach, 2019), Fortune (Eversley, 2019), and Newswise (Flamisch, 2019), among others.

In this application, we focus on the youth voter subset. The data was used to examine the effect of preregistration laws on turnout among young voters (Holbein and Hillygus, 2016). The data set has $n = 44,821$ observations and $p = 15$ variables (Table 1), with sensitive attributes such as Family Income and Voted and pseudo-identifiers such as gender, race information, and location information that can be used by adversaries to identify subjects or to link to other databases. Therefore, it is important to ensure that the private information is protected before releasing the data.

Variable	category (percentage)
Voted	yes (34.7), no (63.3)
Preregistration State	yes (10.5), no (89.5)
Age (years)	18 (20.5), 19 (19.7), 20 (20.0), 21 (20.0), 22 (19.8)
Married	yes (7.5), no (92.5)
Female	yes (50.7), no (40.3)
Family Income[‡]	14 levels (3.1, 5.9, 26.5)[‡]
College Degree	yes (3.1), no (96.9)
White	yes (69.3), no (31.7)
Hispanic	yes (14.6), no (85.4)
Registration Status	yes (52.4), yes (47.8)
Metropolitan Area	yes (77.9), no (22.1)
Length of Residence	<1 month (2.7), 1-6 months (20.0), 7-11 months (6.9), 1-2 years (15.0), 3-4 years (9.8), $\geq$ 5 years (45.6)
Business/Farm Employment	yes (12.6), no (12.62)
In-Person Interview	yes (36.7), no (63.3)
DMV Registration	yes (12.7), no (87.3)

For Family Income, the minimum, medium and maximum percentages among the 14 levels are listed.

Table 1: List of variables and their values from youth voter subset in the 2000-2012 CPS data.

4.2 Implementation

Similar to the simulation studies, we compare the STEPS procedure in data utility against the modified UH approach and PrivBayes. Hay et al. (2016) observe that when n or ϵ is large, it is unlikely that any of the more complex algorithms will beat a simpler and easier-to-deploy flat algorithm. Similar observations regarding the Laplace sanitizer are also obtained in Bowen and Liu (2020), especially when p is large. Given that the voter data set has a large n (44,821) and a relatively large p (15), it will be of interest to see how STEPS fares against the simple flat Laplace sanitizer, which is included as another benchmark.

For STEPS, we applied Algorithm 1. When deciding on the variable order for partitioning and building the tree, we leveraged domain knowledge in pubic policy instead of using

the exponential mechanism. Our domain expert, a public policy researcher at the Urban Institute, suggested that "Voted", "College Degree", and "Preregistration State" are the top three important variables to practitioners in this data set. "Voted" is ranked first, because this is voter data set and "voted" is perhaps the variable that is of interest to most researchers and practitioners who use the data. Also, this attribute presents an important predictor for turnout in future elections (Malchow, 2004). "College Degree" is ranked the second since scholars have consistently noted that highly educated individuals participate in politics more than the average citizen (Campbell et al., 1980; Rosenstone and Hansen, 1993). Some researchers have also concluded that education is the socio-demographic variable most strongly correlated with turnout (Wolfinger et al., 1980), which also relates to education status. The third variable is "Pre-registration State". McDonald and Thornburg (2010) found higher turnout rates in Florida and Hawaii among those who preregistered (registered before they were able to take part in the next election) compared to those who registered after they turned 18. Pre-registrants were 4.7% more likely to vote in the 2008 election than those who registered after they turned 18. For comparison, we built a tree with $L = 2$ (partitioning by "voted" then "College Degree"), and a tree with $L = 3$ (partitioning by "voted" then "College Degree", then "Pre-registration State"). Figure 4 shows the tree structure for STEPS at $L = 2$. Since "family income" has the most categories, thus $b = 14$ and the phantom categories are created for all the other 14 attributes to have $b = 14$ universally. These phantom categories are created solely for the purposes of applying the inconsistency correction rule in Eqns (2) and (3), and will be removed once the synthetic data are generated.

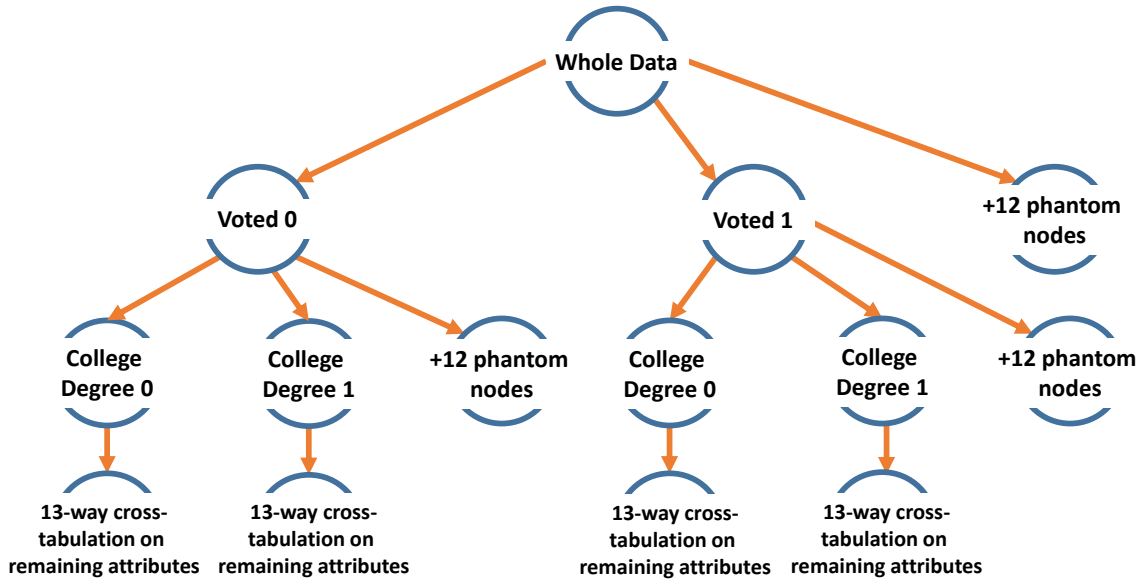


Figure 4: A sketch of the constructed tree in STEPS-2.

For UH, We did not set the number of layers at $L = \log_b N$ per the original procedure Hay et al. (2010), but used $L = 2$ and $L = 3$. The main reason is the computational constraints and practical limitations. If we had used $b = 14$ with 15 attributes, with $L = 15$ and

$N = 14^{15}$, and the computational complexity would be $O(N) = O(14^{15})$. In each layer for $L = 2$ and $L = 3$, we randomly chose an attribute to partition the data, sanitized the node counts, and then applied Eqns (2) and (3) to obtain the final differentially private histogram, from which data were synthesized.

For PrivBayes, we applied the GitHub codes (Ping, 2018). The degree of the Bayesian network was set at 2 and the total ϵ was divided in half between selecting a Bayesian network vs. sanitizing the resultant joint distribution.

The flat Laplace sanitizer injected independent noises from Laplace$(0, \epsilon^{-1})$ to each of the $1,290,240$ cells, after removing the impossible case of a person voting if they are not registered, from the 15-way cross-tabulation (Table 2). Given the enormous number of cells, the majority of the cells are empty (98.35%). Since the empty cells are sample zeros rather than population zeros, they were and should be sanitized rather than being kept at zero. Data were synthesized from the 15-dimensional sanitized table.

Cell size	0	1	2	3	4	5	> 5
Number of Cells	1,268,911	14,061	3,545	1,508	740	425	1,050
Proportion	98.35%	1.09%	0.27%	0.12%	0.06%	0.03%	0.08%

Table 2: Summary of cell sizes and frequencies in the full cross tabulation of the youth voter data.

We varied privacy budget at $\epsilon \in \exp\{-2, -1, 0, 1, 2\}$ to examine how ϵ affects the utility of the synthetic data in each of the DIPS methods. We generated 5 synthetic data sets, each at a budget of $\epsilon/5$ to account for the synthesis and sanitization variability. We ran 10 repetitions to quantify the stability of the DIPS methods.

4.3 Statistical Utility Assessment

We assess the statistical utility of the synthetic data via three analyses: the SPECKS metric, chi-squared tests of association, and a difference-in-differences (DID) model. As discussed in Sec 2.4, the SPECKS metric measures the worst-case separation in two data sets as a whole, whereas the latter two focus on exploring relationships among the attributes that practitioners would be interested in learning from this data set. Interested readers may refer to Bowen and Snoke (2019) for further discussion on different types of utility analysis.

In the SPECKS analysis, we used the logistic regression that contains the main effects and first order interactions among all the attributes as predictor as the classifier. The results from the 5 synthetic data sets were averaged.

For the chi-squared tests of association, we conducted tests for all possible 2-way tables (105 in total) across the 15 attributes to see how well the statistically significant 2-way associations detected in the original data are preserved in the synthetic data. Since the goal here is not to to pick at least one statistically significant associations out of 105 tables, we did not apply the multiplicity adjustment. We first conducted the 105 chi-squared tests and obtained the chi-squared test statistic X^2_{jk} for the k-th test in synthetic data set j for $k = 1, \ldots, 105$ and $j = 1, \ldots, m = 5$. The degree of freedom of the k-th chi-squared test

is denoted by ν_k. We then calculated the combined p-value for the k-th test across the 5 synthetic sets via the procedure in Li et al. (1991). We note the p-value combination rule is developed in the framework of the Wald-type chi-squared test in Li et al. (1991), whereas the chi-squared test in our case is a score test. We don't expect this discrepancy to be a problem given the large sample size in this case study. Specifically, the combined p-value for the k-th test is calculated as

$$\Pr(\hat{D}_k > F_{\nu_k,\eta_k}(1-\alpha)), \tag{11}$$

where $\hat{D}_k = (\overline{X_k^2}\nu_k^{-1} - ((m+1)(m-1)^{-1})\, r_{D,k})(1+r_{D,k})^{-1}$ is the test statistic, $\overline{X_k^2} = m^{-1}\sum_{j=1}^m X_{jk}^2$, $r_{D,k} = (1+m^{-1})\left[(m-1)^{-1}\sum_{j=1}^m \left(\sqrt{X_{jk}^2} - \sqrt{\overline{X_k^2}}\right)\right]$, $\eta_k = \nu_k^{-3/m}(m-1)(1+r_{D,k}^{-1})^2$, and $F_{\nu_k,\eta_k}(1-\alpha)$ is the $(1-\alpha)$ percentile of the F_{ν_k,η_k} distribution.

For the third analysis, we ran the DID logistic regression in Holbein and Hillygus (2016) to examine the effects of "Preregistration State" and 'Registration Status" on "Voted"; all 14 attributes are included as predictors plus an interaction term between "Preregistration State" and "Registration Status". "Age" is treated as a numerical predictor whereas the others are categorical, leading to a total of 32 regression coefficients $\boldsymbol{\beta}$, including the intercept. We first applied the inferential combination rules from Eqns (8) to (10) to obtain the 95% CIs for $\boldsymbol{\beta}$. We then calculated the CI overlap metric (Karr et al., 2006), which is defined as

$$0.5\left(\frac{\min(u_o,u_s) - \max(l_o,l_s)}{u_o - l_o} + \frac{min(u_o,u_s) - max(l_o,l_s)}{u_s - l_s}\right), \tag{12}$$

where u_o, l_o and u_s, l_s are the upper and lower bounds for the original and synthetic CIs respectively. This metric measures how much the CI estimation for a parameter based on the original data overlaps that based on the synthetic data. In other words, a value of 1 corresponds to perfect match between the CIs based on the synthetic and original data; and a value of 0 means there is no overlap between the two CIs (regardless of the degree of non-overlap). When one CI is completely contained within the other, the CI overlap value is greater than 0.5.

4.3.1 Utility Result Summary

We summarize the utility results from all three types of analysis in Table 3. The numbers in parentheses are their performance rank, where 1 is the best and 6 is the worst. The overall score is the sum of all ranks: the smaller, the better. The consistency rate column refers to the percentage that the chi-squared test conclusions are consistent between the original and synthetic data out of 105 tests at the significance levels of $\alpha = \{1, 5, 10\}\%$, respectively. For the CI overlap metric, we report that the average CI overlap values across all 32 coefficients, as well as those on the coefficients associated with covariates "Preregistration State", "Registration Status", and the interaction term between "Preregistration State" and "Registration Status" (Prereg. x Reg.), which are the three covariates are of primary interest in the DID model of Holbein and Hillygus (2016).

log(ϵ)	DIPS Method	SPECKS: KS Distance	Consistency Rate with α			Confidence Interval Overlap				Overall Score
			1%	5%	10%	All Coeff.	Preg. State	Reg. Status	Prereg. x Reg.	
-2	flat Laplace	0.981 (6)	0.190 (1)	0.210 (1)	0.210 (1)	0.092 (3)	0.872 (3)	0.637 (3)	0.872 (3)	21
-2	PrivBayes	0.462 (1)	0.152 (6)	0.143 (6)	0.105 (6)	0.043 (6)	0.681 (6)	0.000 (6)	0.681 (6)	43
-2	UH modified-2	0.973 (4)	0.162 (2)	0.152 (2)	0.114 (2)	0.099 (1)	0.865 (4)	0.602 (4)	0.865 (4)	23
-2	UH modified-3	0.957 (3)	0.162 (2)	0.152 (2)	0.114 (2)	0.093 (2)	0.848 (5)	0.495 (5)	0.848 (5)	26
-2	STEPS-2	0.975 (5)	0.162 (2)	0.152 (2)	0.114 (2)	0.090 (4)	0.874 (1)	0.661 (1)	0.874 (1)	18
-2	STEPS-3	0.956 (2)	0.162 (2)	0.152 (2)	0.114 (2)	0.090 (4)	0.873 (2)	0.659 (2)	0.873 (2)	18
-1	flat Laplace	0.978 (6)	0.181 (1)	0.210 (1)	0.210 (1)	0.102 (1)	0.871 (3)	0.638 (3)	0.871 (3)	19
-1	PrivBayes	0.422 (1)	0.152 (6)	0.143 (6)	0.105 (6)	0.050 (6)	0.681 (6)	0.000 (6)	0.681 (6)	43
-1	UH modified-2	0.971 (4)	0.162 (2)	0.152 (4)	0.114 (4)	0.098 (2)	0.865 (4)	0.602 (4)	0.865 (4)	28
-1	UH modified-3	0.954 (2)	0.162 (2)	0.152 (4)	0.114 (4)	0.092 (3)	0.839 (5)	0.445 (5)	0.839 (5)	30
-1	STEPS-2	0.971 (4)	0.162 (2)	0.171 (2)	0.133 (2)	0.090 (5)	0.874 (1)	0.667 (1)	0.874 (1)	18
-1	STEPS-3	0.954 (2)	0.162 (2)	0.171 (2)	0.133 (2)	0.091 (4)	0.874 (1)	0.666 (2)	0.874 (1)	16
0	flat Laplace	0.968 (6)	0.190 (1)	0.210 (1)	0.210 (1)	0.108 (1)	0.871 (4)	0.641 (3)	0.871 (4)	21
0	PrivBayes	0.272 (1)	0.152 (6)	0.143 (6)	0.114 (4)	0.076 (6)	0.943 (1)	0.000 (6)	0.962 (1)	31
0	UH modified-2	0.964 (4)	0.162 (4)	0.152 (4)	0.114 (4)	0.106 (2)	0.864 (5)	0.605 (4)	0.864 (5)	32
0	UH modified-3	0.945 (2)	0.162 (4)	0.152 (4)	0.114 (4)	0.090 (3)	0.838 (6)	0.446 (5)	0.838 (6)	34
0	STEPS-2	0.965 (5)	0.181 (2)	0.171 (2)	0.133 (2)	0.090 (3)	0.873 (2)	0.671 (1)	0.873 (2)	19
0	STEPS-3	0.945 (2)	0.181 (2)	0.171 (2)	0.133 (2)	0.090 (3)	0.873 (2)	0.671 (1)	0.873 (2)	16
1	flat Laplace	0.930 (4)	0.276 (1)	0.371 (1)	0.448 (1)	0.092 (4)	0.867 (4)	0.651 (4)	0.867 (4)	23
1	PrivBayes	0.187 (1)	0.171 (4)	0.171 (6)	0.133 (6)	0.102 (2)	0.978 (1)	0.847 (1)	0.990 (1)	22
1	UH modified-2	0.944 (5)	0.162 (5)	0.200 (4)	0.219 (4)	0.125 (1)	0.863 (5)	0.612 (5)	0.863 (5)	34
1	UH modified-3	0.919 (2)	0.162 (5)	0.181 (5)	0.219 (4)	0.096 (3)	0.837 (6)	0.453 (6)	0.837 (6)	37
1	STEPS-2	0.944 (5)	0.200 (2)	0.257 (2)	0.238 (2)	0.091 (5)	0.871 (2)	0.677 (2)	0.871 (2)	22
1	STEPS-3	0.919 (2)	0.200 (2)	0.248 (2)	0.238 (2)	0.084 (6)	0.871 (2)	0.676 (3)	0.871 (2)	21
2	flat Laplace	0.833 (2)	0.467 (1)	0.552 (1)	0.629 (1)	0.104 (1)	0.862 (4)	0.684 (4)	0.862 (4)	18
2	PrivBayes	0.157 (1)	0.190 (6)	0.181 (6)	0.143 (7)	0.099 (3)	0.988 (1)	0.983 (1)	0.971 (1)	26
2	UH modified-2	0.885 (5)	0.352 (4)	0.400 (4)	0.410 (6)	0.100 (2)	0.860 (5)	0.635 (5)	0.860 (5)	36
2	UH modified-3	0.843 (3)	0.305 (5)	0.390 (5)	0.457 (4)	0.076 (5)	0.849 (6)	0.553 (6)	0.849 (6)	40
2	STEPS-2	0.886 (6)	0.400 (2)	0.448 (2)	0.486 (2)	0.093 (4)	0.865 (2)	0.696 (2)	0.865 (2)	22
2	STEPS-3	0.843 (3)	0.400 (2)	0.448 (2)	0.486 (2)	0.076 (5)	0.865 (2)	0.695 (3)	0.865 (2)	21

Table 3: Summary utility analysis. The numbers in parentheses are the performance ranks with 1 being the best and 6 being the worst. Overall score is the sum of all ranks: the smaller, the better.

Overall, the utility in each analysis increases with ϵ for each DIP method, with the only exception on the CI overlap values from the DID model, where the improvement with ϵ is not obvious. All analyses taken together, STEPS-2 and STEPS-3 with $L = 2$ and $L = 3$ are the best, followed by the flat Laplace sanitizer. PrivBayes performs the best with regard to the SPECKS analysis, by a large margin, compared to the rest, but does not do well in the chi-squared tests or the DID model CI overlap analysis. The Laplace sanitizer performs the best in the chi-squared tests, slightly worse than STEPS on the other analysis especially when ϵ is small. The STEPS method outperforms the random partitioning approach via the modified UH method on almost all metrics, demonstrating the advantages of informed partitioning.

In what follows, we provide more details on the results in each of the utility analyses.

4.3.2 SPECKS Analysis for General Utility

Figure 5 depicts the results on the SPECKS analysis (the standard deviations on the KS distance across the repeats are two orders of magnitude smaller than the average KS distance and barely visible on the plots). For all methods, the KS distance decreases as ϵ increases, as expected, providing empirical evidence that SPECKS does what it is supposed to measure. Overall, PrivBayes outperforms all of the other methods for all levels of ϵ. STEPS-2 and STEPS-3 are very similar and slightly outperform the flat Laplace sanitizer (until $\epsilon = e^2$) and the UH modified.

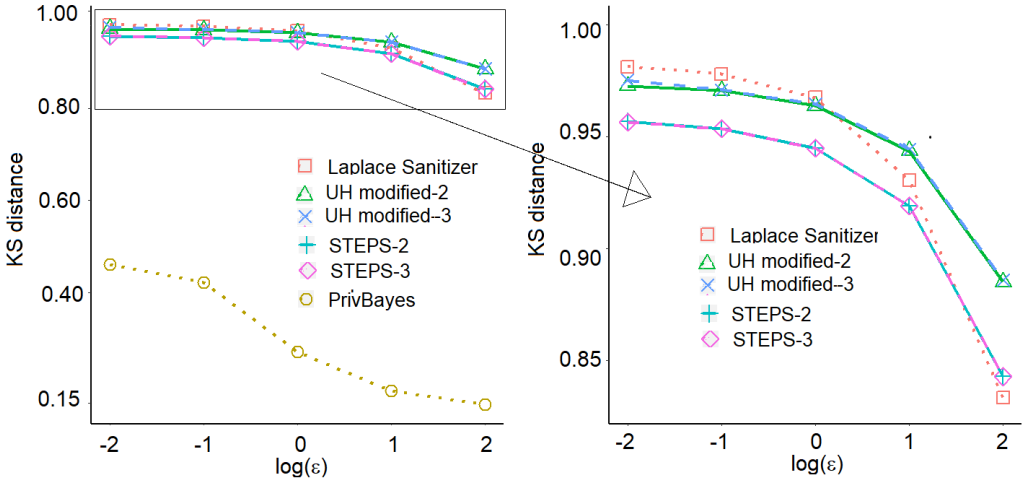


Figure 5: SPECKS analysis. The plot on the left presents all the methods and the one on the right is a zoom-in on the boxed area.

4.3.3 Chi-squared Tests of Association

Figure 6 presents the chi-squared test statistical significance consistency rates at $\alpha = 1\%$, 5%, 10%, respectively. For all values of ϵ, the Laplace sanitizer performs the best, followed by the STEPS methods. The UH modified methods perform slightly worse than the STEPS methods while PrivBayes remains at consistently low rates across all levels of ϵ. We also present in the Appendix Tables 4 to 9 that contain the cross-tabulations of the original vs. sanitized p-value by categories ≤ 0.01, $(0.01, 0.05]$, $(0.05, 0.1]$, and > 0.1. The tables suggest there is improvement in the alignment between the original and sanitized p-values as ϵ increases, but in a less satisfactory manner with this more granular categorization, even for ϵ as large as $e^2 \approx 7.4$, compared to the binary classification on the p-values plotted in Figure 6.

4.3.4 Difference-in-Differences (DID) Model

Figure 7 shows the results for the average CIs across all 32 regression coefficients, that on the coefficients for "Preregistration State", "Registration Status", and the interaction term between "Preregistration State" and "Registration Status" (Prereg. x Reg.), respectively

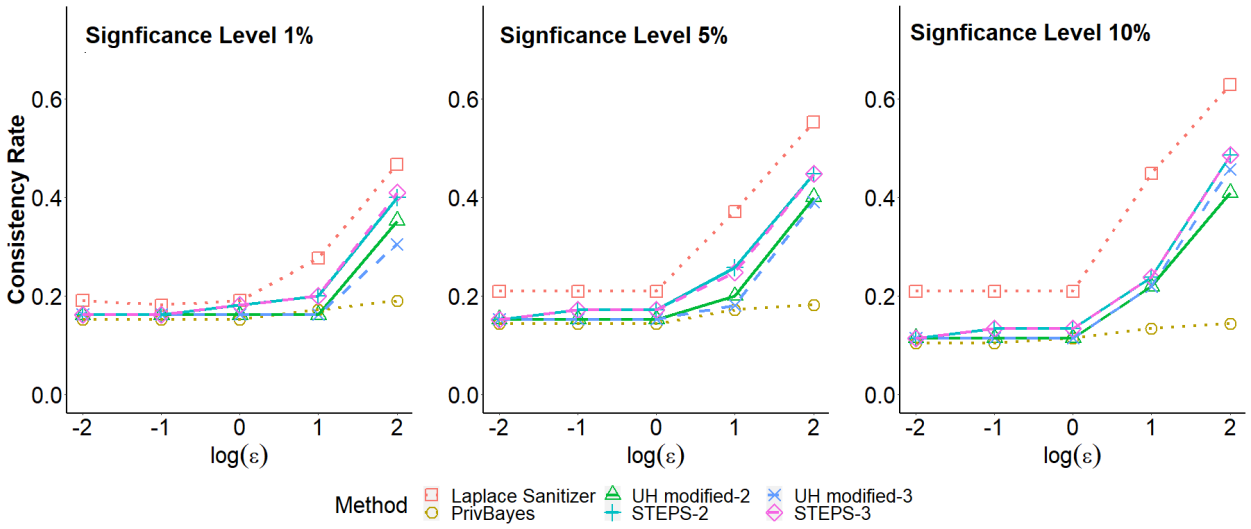


Figure 6: Consistency rate on the statistical significance in the 105 Chi-square tests of association based on the original and the synthetic data.

form the DID model. The effects of the latter three covariates are of primary interest in Holbein and Hillygus (2016).

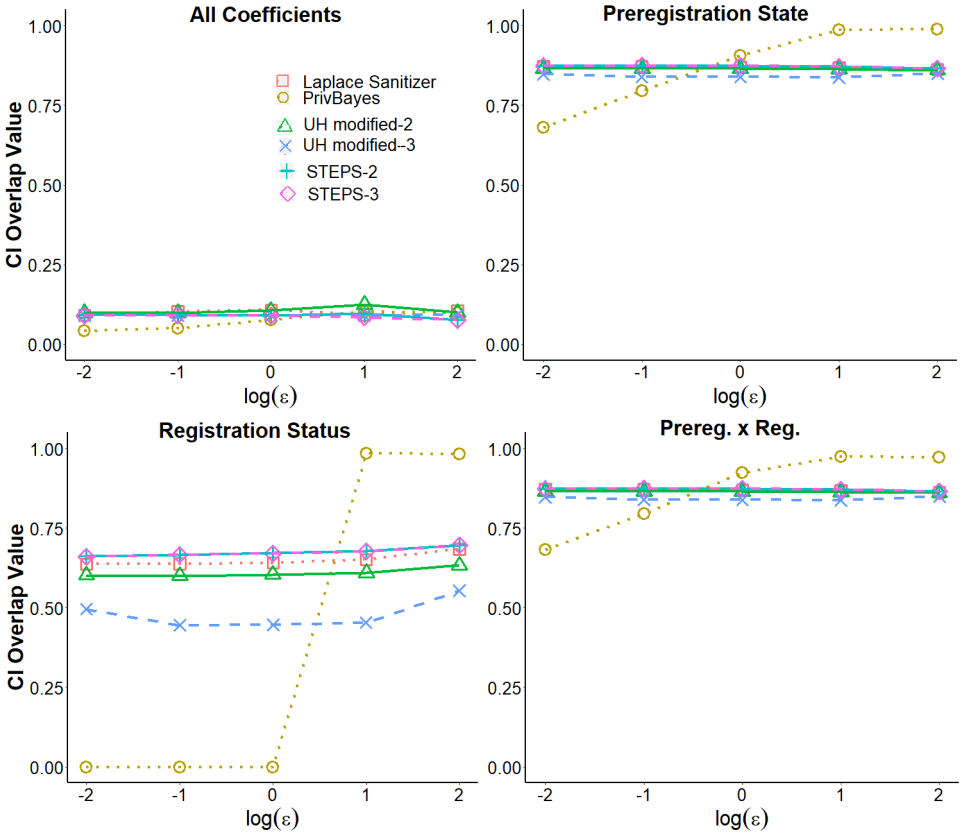


Figure 7: Confidence interval (CI) overlap value analysis averaged across the 32 regression coefficients from the DID model fitted to the synthetic data generated from the flat Laplace sanitizer, PrivBayes, STEPS, and a modified UH with random partition.

For the average CI overlap value across the 32 coefficients, since a non-trivial proportion of the CI overlap values are close 0, the averaged CI is close to 0 and is not very informative in differentiating the methods. For the three covariates of interest, the CI overlap values are around 0.75 or larger in most cases, suggesting good agreement between the CIs based on the synthetic and original data. Among the four DIPS methods, PrivBayes has the lowest CI overlap values when $\log(\epsilon) < 0$, and performs better as ϵ increase. The CI overlap values are similar for all the other methods and across the examined ϵ values.

5 Concluding Remarks

We propose the STEPS procedure to synthesize differentially private individual-level data. We also develop a propensity-based general utility metric, SPECKS, for assessing the similarity between the original and synthetic data. The STEPS method leverages the inherent information in the data or public domain knowledge to build a hierarchical tree among the attributes. The sanitized counts of the nodes closer to the root of the tree (low-order marginals) have smaller mean squared errors than the nodes further away from the root because the former are weighted averages of multiple sources of information. STEPS therefore preserves the information on more important attributes better than on less important ones, placing those attributes closer to the root of the tree.

We used the PrivBayes codes on GitHub at Ping (2018) developed by Ping and Stoyanovich (2017). We noticed during the implementation that we needed to set a random seed in each synthetic generate; otherwise, every synthetic set is the same for a given ϵ. Users of the codes need to keep that in mind when generating multiple synthetic sets to measure the synthesis uncertainty. This is just one example on the importance of good documentation when publishing and sharing nodes for implementing differentially private mechanisms and data synthesis methods to avoid misleading interpretation or results. Kifer et al. (2020) encourage that every DP mechanism provides an accompanying mathematical proof on DP (not a just reference to the literature) along with proper code review and computational-aided verification/testing tools. We will publish the codes used in the simulation and cases studies in this paper on GitHub to facilitate the practical implementation of the proposed method and for reproducibility checks.

For future work, we will look into developing general recommendations on the privacy budget allocation scheme between building a tree and sanitizing node counts among the tree layers. In addition, we will investigate the feasibility of developing a stopping rule on the tree height, from the utility and computational prospective, instead of pre-specifying the number of layers. We also plan to compare SPECKS with other propensity-score based general utility measures via comprehensive empirical studies. A review commented on the CI overlapping metric (Karr et al., 2006) we employed for the case study, and questioned whether the degree of non-overlapping between two non-overlapping CIs should be take into account. The existing metric set the metric at 0 for all non-overlapping CI cases, regardless of the distance between two CIs. This is an interesting question that may lead to a new CI overlap metric. Our initial thoughts are as follows. If negative values are to be assigned to non-

overlapping CIs, the values should consider both the distance (say, measured by the difference of the lower bound of the CI located on the right and the upper bound of the CI located on the left), the width of the CIs, whether the original and synthetic CI widths should be weighted differently, and be easy to interpret.

Acknowledgement

We thank Madeline Brown, a Policy Assistant, at the Urban Institute for her expert knowledge in selecting variable order as a domain expert in the case study. We also thank the associate editor and two referees who provided useful feedback to improve the manuscript.

We also thank the Editor, the Associate Editor and the two referees for their valuable comments and suggestions that improved the quality of the manuscript.

References

J. M. Abowd and I. M. Schmutte. Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods, 2015. URL http://doi.org/10.5281/zenodo.345385.

A. ABRAMS. Voter turnout surged among people with disabilities last year. activists want to make sure that continues in 2020, 2019. URL https://time.com/5622652/disability-voter-turnout-2020/.

C. M. Bowen and F. Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307, 2020.

C. M. Bowen and J. Snoke. Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge. *arXiv preprint arXiv:1911.12704*, 2019.

A. Campbell, P. E. Converse, W. E. Miller, and D. E. Stokes. *The american voter*. University of Chicago Press, 1980.

R. Cummings and D. Durfee. Individual sensitivity preprocessing for data privacy. *arXiv preprint arXiv:1804.08645*, 2018.

B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: Optimizing noise sources and consistency. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2011)*, pages 217–228. ACM SIGMOD, 2011.

J. Drechsler. *Synthetic datasets for Statistical Disclosure Control*. Springer, New York, 2011.

C. Dwork. Differential privacy: A survey of results. In *TAMC 2008, LNCS 4978*, pages 1–19. Springer-Verlag Berlin Heidelberg, 2008.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

C. Dwork and G. N. Rothblum. Concentrated differential priacy. *arXiv preprint arXiv:1603.01887*, 2016.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.

E. C. Eugenio and F. Liu. Cipher: Construction of differentially private microdata from low-dimensional histograms via solving linear equations with tikhonov regularization. *arXiv preprint*, (1812.05671), 2018.

M. Eversley. Black women voters will be central to the 2020 presidential election, experts predict, 2019. URL https://fortune.com/2019/06/20/black-women-voters-2020-election/.

S. Flamisch. Voter turnout surging among people with disabilitie, 2019. URL https://www.newswise.com/articles/voter-turnout-surging-among-people-with-disabilities.

J. Gardner, L. Xiong, Y. Xiao, J. Gao, A. R. Post, X. Jiang, and L. Ohno-Machado. Share: system design and case studies for statistical health information release. *Journal of the American Medical Informatics Association*, 20(1):109–116, 2013.

M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032, 2010.

M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang. Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the 2016 International Conference on Management of Data*, pages 139–154. ACM, 2016.

C. Hill and J. Grumbach. An excitingly simple solution to youth turnout, for the primaries and beyond, 2019. URL https://www.nytimes.com/2019/06/26/opinion/graphics-an-excitingly-simple-solution-to-youth-turnout-for-the-primaries-and-beyond.html.

J. B. Holbein and D. S. Hillygus. Making young voters: The impact of preregistration on youth turnout. *American Journal of Political Science*, 60(2):364–382, 2016. ISSN 1540-5907.

A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232, 2006.

D. Kifer, S. Messing, A. Roth, A. Thakurta, and D. Zhang. Guidelines for implementing and auditing differentially private systems. *arXiv preprint arXiv:2002.04049*, 2020.

C. Li, M. Hay, G. Miklau, and Y. Wang. A data- and workload-aware algorithm for range queries under differential privacy. *Proceedings of International Conference on Very Large Data Bases (PVLDB)*, 7:341–352, 2014a.

H. Li, L. Xiong, and X. Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Proc. 17th International Conference on Extending Database Technology (EDBT)*, pages 475–486, 2014b.

H. Li, L. Xiong, Z. Ji, and X. Jiang. Partitioning-based mechanisms under personalized differential privacy. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 615–627. Springer, 2017.

K.-H. Li, X.-L. Meng, T. E. Raghunathan, and D. B. Rubin. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, pages 65–92, 1991.

R. Little. Statistical analysis of masked data. *Journal of the Official Statistics*, 9:407, 1993.

R. Little, F. Liu, and T. Raghunathan. Statistical disclosure techniques based on multiple imputation. In A. Gelman and X.-L. Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An essential journey with Donald Rubin's statistical family*, page Chapter II.13. John Wiley & Sons, 2004.

F. Liu. Model-based differentially private data synthesis. *arXiv preprint arXiv:1606.08052*, 2016.

F. Liu. Generalized gaussian mechanism for differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):747 – 756, 2019.

F. Liu and R. Little. Smike vs. data swapping and pram for statistical disclosure limitation in microdata: A simulation study. *Proceedings of 2003 American Statistical Association Joint Statistical Meeting*, 2003.

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. *IEEE ICDE IEEE 24th International Conference*, pages 277 – 286, 2008.

H. Malchow. Predicting turnout in a presidential election. *CAMPAIGNS AND ELECTIONS*, 25(8):38–41, 2004.

M. P. McDonald and M. Thornburg. Registering the youth through voter preregistration. *NYUJ Legis. & Pub. Pol'y*, 13:551, 2010.

F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30. ACM, 2009.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

H. Ping. Datasynthesizer. https://github.com/DataResponsibly/DataSynthesizer (accessed online 3/14/2020), 2018.

H. Ping and J. Stoyanovich. Datasynthesizer: Privacy-preserving synthetic datasets. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017.

H. Ping, J. Stoyanovich, and B. Howe. Datasynthesizer: Privacy-preserving synthetic datasets. *SSDBM '17: Proceedings of the 29th International Conference on Scientific and Statistical Database Management June 2017*, (42):1–5, 2017.

W. Qardaji, W. Yang, and N. Li. Differentially private grids for geospatial data. *Proceedings of 2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 757–768, 2013a.

W. Qardaji, W. Yang, and N. Li. Understanding hierarchical methods for differentially private histograms. *Proceedings of the VLDB Endowment*, 6(14):1954–1965, 2013b.

T. E. Raghunathan, J. P. Reiter, and D. B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official Statistics*, 19(1):1–16, 2003.

J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

J. P. Reiter. Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review*, 77(2):179–195, 2009.

S. J. Rosenstone and J. Hansen. *Mobilization, participation, and democracy in America*. Macmillan Publishing Company,, 1993.

D. B. Rubin. Discussion statistical disclosure limitation. *Journal of official Statistics*, 9(2): 461, 1993.

J. W. Sakshaug and T. E. Raghunathan. Synthetic data for small area estimation. In *International Conference on Privacy in Statistical Databases*, pages 162–173. Springer, 2010.

J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663–688, 2018.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

R. E. Wolfinger, S. J. Rosenstone, and S. J. Rosenstone. *Who votes?*, volume 22. Yale University Press, 1980.

M.-J. Woo, J. P. Reiter, A. Oganian, and A. F. Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. *Secure Data Management*, 6358:150–168, 2010.

Y. Xiao, J. Gardner, and L. Xiong. Dpcube: Releasing differentially private data cubes for health information. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1305–1308. IEEE, 2012.

J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS) - Invited Paper from SIGMOD 2016*, 24, 2017.

Appendix

$\log(\epsilon)$	sanitized					consistency count‡
	Original	< 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	
-2	< 0.01	5	5	0	79	15
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	2	0	2	
	> 0.1	1	0	0	10	
-1	< 0.01	4	6	0	79	14
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	2	0	2	
	> 0.1	1	0	0	10	
0	< 0.01	5	5	0	79	15
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	2	0	2	
	> 0.1	1	0	0	10	
1	< 0.01	19	12	4	54	27
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	3	1	0	0	
	> 0.1	3	0	0	8	
2	< 0.01	41	11	6	31	45
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	4	0	0	0	
	> 0.1	4	1	2	4	

‡ sum of diagonal counts.

Table 4: The cross-tabulation of the 105 p-values by categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the Laplace sanitizer.*

$\log(\epsilon)$	Original	sanitized < 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	consistency count[‡]
-2	< 0.01	0	0	0	89	11
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
-1	< 0.01	0	0	0	89	11
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
0	< 0.01	0	0	1	88	11
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
1	< 0.01	2	1	0	86	13
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
2	< 0.01	4	0	0	85	15
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	

[‡] sum of diagonal counts.

Table 5: The cross-tabulation of the 105 p-values in categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the PrivBayes.*

$\log(\epsilon)$	sanitized					consistency count[‡]
	Original	< 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	
-2	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
-1	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
0	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
1	< 0.01	1	7	4	77	10
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	2	0	2	
	> 0.1	0	0	2	9	
2	< 0.01	22	8	5	54	29
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	1	3	
	> 0.1	1	2	1	7	

[‡] sum of diagonal counts.

Table 6: The cross-tabulation of the 105 p-values by categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the UH Modified-2.*

$\log(\epsilon)$	Original	sanitized < 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	consistency count[‡]
-2	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
-1	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
0	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
1	< 0.01	1	6	5	77	10
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	2	0	2	
	> 0.1	0	1	1	9	
2	< 0.01	20	13	8	48	25
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	1	1	0	2	
	> 0.1	3	2	1	5	

[‡] sum of diagonal counts.

Table 7: The cross-tabulation of the 105 p-values by categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the UH Modified-3.*

log(ϵ)	sanitized					consistency count[‡]
	Original	< 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	
-2	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
-1	< 0.01	1	2	0	86	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
0	< 0.01	3	0	0	86	14
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
1	< 0.01	5	8	4	72	13
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	1	2	8	
2	< 0.01	30	7	8	44	36
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	4	1	0	6	

[‡] sum of diagonal counts.

Table 8: The cross-tabulation of the 105 p-values by categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the STEPS-2.*

$\log(\epsilon)$	Original	sanitized < 0.01	$0.01 - 0.05$	$0.05 - 0.1$	> 0.1	consistency count[‡]
-2	< 0.01	1	0	0	88	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
-1	< 0.01	1	2	0	86	12
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
0	< 0.01	3	0	0	86	14
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	0	0	11	
1	< 0.01	5	8	3	73	14
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	0	2	0	9	
2	< 0.01	30	7	8	44	36
	$0.01 - 0.05$	0	0	0	1	
	$0.05 - 0.1$	0	0	0	4	
	> 0.1	3	2	0	6	

[‡] sum of diagonal counts.

Table 9: The cross-tabulation on the 105 p-values by categories < 0.01, $0.01 - 0.05$, $0.05 - 0.1$, *and* > 0.1 *for the original and DIPS data generated by the STEPS-3.*