Journal of Privacy and Confidentiality
Vol. 16 (1) 2026

Submitted February 2025
Published April 2026

MAXIMIZING UTILITY FOR VECTOR-WEIGHTED PSEUDO POSTERIOR MECHANISMS UNDER DIFFERENTIAL PRIVACY

JINGCHEN HU, TERRANCE D. SAVITSKY, AND MATTHEW R. WILLIAMS

Binghmaton University, Mathematics and Statistics Department, 4400 Vestal Parkway East, Binghamton, NY 13902
e-mail address: mhu7@binghamton.edu

U.S. Bureau of Labor Statistics, Office of Survey Methods Research, Suite 5930, 2 Massachusetts Ave NE Washington, DC 20212
e-mail address: Savitsky.Terrance@bls.gov

RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709
e-mail address: mrwilliams@rti.org

Abstract. The risk-weighted pseudo posterior mechanism provides a practical framework for privacy protection that takes advantage of the availability of posterior sampling approaches, creating a synthesizer for microdata dissemination. The flexibility of the approach lies in the user-specification of the individualized risks and the mapping of risks to weights. However, this raises the question of which weighting approach is optimal. In this work, we develop a recursive approach to algorithmically induce an optimal weighting strategy given an initial suboptimal strategy. This "re-weighting" strategy applies to any vector-weighted pseudo posterior mechanism under which a vector of observation-indexed weights are used to downweight likelihood contributions for high disclosure risk records. We demonstrate our method on two different vector-weighted schemes that target high-risk records (one close to optimal and one not). Our new method for constructing record-indexed downweighting maximizes the data utility under any privacy budget for the vector-weighted synthesizers by adjusting the by-record weights, such that their individual risk contributions (e.g. Lipschitz bounds) approach the risk bound for the entire database. Our method achieves an $\epsilon-$asymptotic differential privacy (aDP) guarantee, globally, over the space of databases. We illustrate our methods using simulated highly skewed count data. We also apply our methods to a sample of the Survey of Doctorate Recipients and demonstrate the practicality of our methods.

1. Introduction

Publishing survey and census data equipped with a privacy guarantee to limit the risk of respondent reidentification is an important goal for government statistical agencies and private companies, worldwide. A commonly-used approach to encode privacy protection into data released to the public generates synthetic data from statistical models estimated

Key words and phrases: data privacy protection, differential privacy, microdata dissemination, pseudo posterior mechanism, synthetic data, utility maximization, vector-weighted.

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.979

on confidential, private data for proposed release by statistical agencies (11; 9). This data synthesis approach replaces the confidential database with multiple synthetically generated record-level databases. When released to the public, such synthetic databases are used in any analyses as if the synthetic database comprised the real, confidential record-level data. Publication of the synthetic databases encoded with privacy protection replaces multiple queries performed on a summary statistic, making this approach independent of the specific queries performed by users or putative intruders.

We focus on a class of synthetic data generating mechanisms that are differentially private. We next briefly review the definition of differential privacy and the related exponential mechanism and pseudo-posterior mechanism on which we will further optimize in order to gain additional utility under the same privacy protection level ϵ. The key challenge is finding an *optimal mapping* of individual level risk to weights used for adjusting the mechanism via a weighted loss function. Too many individuals with weights close to 1 will lead to unacceptably high risk, while too many weights close to 0 will effectively remove individuals completely and degrade utility.

1.1. **Differential privacy.**

Our focus metric for measuring the relative privacy guarantee of our pseudo posterior synthesizing data mechanism introduced in the sequel is differential privacy (3). We next provide a definition for differential privacy (10).

Definition 1.1 (Differential Privacy)**.** *Let* $\mathbf{x}$ *be a database in input space* $\mathcal{X}^n$*, where* $\mathcal{X}^n$ *denotes a space of databases of size (number of observations)* n*. Let* $\mathcal{M}$ *be a randomized mechanism such that* $\mathcal{M}() : \mathcal{X}^n \rightarrow O$*. Then* $\mathcal{M}$ *is* $\epsilon-$*differentially private if*

$$\frac{Pr[\mathcal{M}(\mathbf{x}) \in O]}{Pr[\mathcal{M}(\mathbf{y}) \in O]} \leq \exp(\epsilon),$$

for all possible outputs $O = Range(\mathcal{M})$ *under all possible pairs of datasets* $\mathbf{x} \in \mathcal{X}^n$ *where* $\mathbf{y} \in \mathcal{X}^{n-1}$ *differs from* $\mathbf{x}$ *by deleting one record or datum (under a leave-one-out (LOO) distance definition).*

Differential privacy is a property of the mechanism or data generating process. A mechanism that meets the definition above is guaranteed to be $\epsilon-$differentially private, or $\epsilon-$DP. Differential privacy is called a "formal" privacy guarantee because the ϵ level or guarantee is independent of the behavior of a putative intruder seeking to re-identify the data and the guarantee is not lessened by the existence of other data sources that may contain information about the same respondents included in $\mathcal{X}^n$.

Differential privacy assigns a disclosure risk for a statistic to be released to the public (e.g., total employment for a state-industry). Call this statistic $g(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}^n$ with the global sensitivity, $\Delta = \sup_{\mathbf{x} \in \mathcal{X}^n} \sup_{\mathbf{y} \in \mathcal{X}^{n-1}:\ \delta(\mathbf{x},\mathbf{y})=1} |g(\mathbf{x}) - g(\mathbf{y})|$, over the space of databases, $\mathcal{X}^n$, where $\delta(\mathbf{x}, \mathbf{y})$ denotes the number of records omitted from $\mathbf{x}$ in database $\mathbf{y} \in \mathcal{X}^{n-1}$. The distance metric, $\delta(\mathbf{x}, \mathbf{y})$ denotes the LOO distance such that $\mathbf{x}$ differs from $\mathbf{y}$ by a single record. Other common definitions of differential privacy use the Hamming-1 distance which defines neighboring databases of the same size ($\mathbf{y} \in \mathcal{X}^n$). In the case of count based statistics of binary data records, there is a direct correspondence between LOO and Hamming-1. If the value of the statistic, g, expresses a high magnitude change after the deletion of a data record in $\mathbf{y}$, then the mechanism will be required to induce a relatively higher level of distortion to g. The more sensitive is a statistic to the change of a record, the higher its disclosure risk.

1.2. **Exponential mechanism for data synthesis.** Our focus in this paper is where the mechanism, $\mathcal{M}$, is a model parameterized by θ from which replicate data are synthesized under an $\epsilon-$DP guarantee. In particular, we leverage the smoothing property of θ to reduce the sensitivity and achieve a privacy guarantee in lieu of added noise to the confidential data distribution. A common approach for generating parameter draws of θ under the statistical model for synthesizing data is the exponential mechanism (EM) of **(author?)** (10), which inputs a non-private deterministic utility function for selecting θ and generates θ in such a way that induces an $\epsilon-$DP guarantee on the overall mechanism. The EM is conditioned on the availability of a global sensitivity over the space of databases, Δ_u for some bounded utility function, $u(\mathbf{x}, \theta)$, defined on the space of databases and the space of parameters, globally.

Definition 1.2. *(Exponential Mechanism) The exponential mechanism releases values of θ from a distribution proportional to,*

$$\exp\left(u(\mathbf{x}, \theta)\right) \xi\left(\theta\right), \tag{1.1}$$

where $u(\mathbf{x}, \theta)$ is a utility function. Let $\Delta_u = \sup_{\mathbf{x}\in\mathcal{X}^n} \sup_{\mathbf{y}:\delta(\mathbf{x},\mathbf{y})=1} \sup_{\theta\in\Theta} |u(\mathbf{x}, \theta) - u(\mathbf{y}, \theta)|$ be the sensitivity, defined globally over $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, the $\sigma-$algebra of datasets, $\mathbf{x}$, governed by product measure, P_{θ_0} and the LOO distance metric, $\delta(\mathbf{x}, \mathbf{y}) = 1$. Then each draw of θ from the exponential mechanism is guaranteed to be $\epsilon = 2\Delta_u-$DP.

In order to set an arbitrary $\epsilon \neq 2\Delta_u$, we must modify the utility function $u(\mathbf{x}, \theta)$. The statistical agency owning the confidential data will typically desire to determine ϵ as a matter of policy and not leave it to be the "natural" $\epsilon = 2\Delta_u$. The simplest and most common approach is to rescale it: $u^*(\mathbf{x}, \theta) = \frac{\epsilon}{2\Delta_u} u(\mathbf{x}, \theta)$ (See 10; 3, among many others).

The EM requires the availability of the sensitivity Δ_u for a chosen utility function $u(\mathbf{x}, \theta)$. **(author?)** (16) and **(author?)** (13) construct utility functions that are naturally bounded over all $\mathbf{x} \in \mathcal{X}^n$; however, they are not generally applicable to any population model and in the latter case are very difficult to implement in a computationally tractable manner since the EM distribution must be sampled by an inefficient random-walk Metropolis-Hastings scheme.

For a Bayesian model utilizing the data log-likelihood as the utility function of the EM, **(author?)** (15) sketch and **(author?)** (12) further illustrate that the EM mechanism becomes the model posterior distribution, which provides a straightforward mechanism from which to draw samples. **(author?)** (2) define a model-based sensitivity, $\sup_{\mathbf{x}\in\mathcal{X}^n} \sup_{\mathbf{y}:\ \delta(\mathbf{x},\mathbf{y})=1} \sup_{\theta\in\Theta} |f_\theta(\mathbf{x}) - f_\theta(\mathbf{y})| \leq \Delta$ that is constructed as a Lipschitz bound. They demonstrate a connection between the Lipschitz bound, Δ and $\epsilon \leq 2\Delta$ for each draw of parameters, θ, where $f_\theta(\mathbf{x})$ denotes the model log-likelihood indexed by θ. The guarantee applies to all databases $\mathbf{x}$, in the space of databases of size n, $\mathcal{X}^n$.

However, computing a finite $\Delta < \infty$ in practice, as acknowledged by **(author?)** (2), is difficult-to-impossible for an unbounded parameter space (e.g. a normal distribution) under simple models, which requires truncation of the parameter space to achieve a finite Δ. This truncation only works for some models to achieve a finite Δ. Moreover, parameter truncation becomes intractable for practical models that utilize a multidimensional parameter space.

1.3. **Pseudo posterior mechanism for data synthesis.** To guarantee the achievement of a finite $\Delta < \infty$ for any synthesizing model over an unbounded parameter space, **(author?)**

(12) propose the *pseudo* posterior mechanism that uses a log-pseudo likelihood with a vector of weights $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n) \in [0,1]^n$ where each α_i exponentiates the likelihood contribution, $p(x_i \mid \theta)$, for each record $i \in (1, \ldots, n)$. Each weight, $\alpha_i \in [0,1]$ is set to be inversely proportional to a measure of *individualized* disclosure risk for record, i, such that the model used to generate synthetic data will be less influenced by relatively high-risk records.

The pseudo posterior mechanism of **(author?)** (12) is formulated as

$$\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \propto \prod_{i=1}^{n} p(x_i \mid \theta)^{\alpha_i} \times \xi(\theta), \tag{1.2}$$

where the $\alpha_i \in [0,1]$ serve to downweight the likelihood contributions such that highly risky records are more strongly downweighted. We must choose some decreasing function $m() : r_i \mapsto [0,1]$ which maps the risk measure $r_i = \sup_{\theta \in \Theta} |f_\theta(x_i)|$ to a set of weights in $[0,1]$, where $f_\theta(x_i) = \log p(x_i \mid \theta)$ denotes the model log-likelihood. We note that $m(0) = 1$ and $m(\infty) = 0$. A key challenge we address in this work is obtaining an *optimal* choice of $m()$.

The differential downweighting of each record intends to better preserve utility by focusing the downweighting on high-risk records. High-risk records tend to be those located in the tails of the distribution where the likelihood is close to 0, and thus the log-likelihood is a large negative number, leading $|f_\theta(x_i)|$ to be high. Downweighting these records allows the preservation of the high mass portions of the data distribution in the generated synthetic data. The method sets $\alpha_i = 0$ for any record with a *non-finite* log-likelihood, which *ensures* a finite $\Delta_{\boldsymbol{\alpha}} = \sup_{\mathbf{x} \in \mathcal{X}^n} \sup_{\mathbf{y}:\ \delta(\mathbf{x},\mathbf{y})=1} \sup_{\theta \in \Theta} |f_\theta^{\alpha}(\mathbf{x}) - \ f_\theta^{\alpha}(\mathbf{y})| < \infty$. We see that $\Delta_{\boldsymbol{\alpha}} \leq \Delta$ since $\alpha_i \leq 1$. Less optimal choices of $m()$ will lead to too many records with weights near 0 (poor utility) or too many records with weights near 1 (bounded but higher $\Delta_{\boldsymbol{\alpha}}$). An optimal $m()$ will yield $\alpha_i < 1$ for only as many records i as needed to achieve a given $\Delta_{\boldsymbol{\alpha}}$.

Definition 1.3. *(Differential Privacy for the Pseudo Posterior Mechanism) The $\boldsymbol{\alpha}-$weighted pseudo synthesizer, $\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x})$, is a privacy mechanism defined in Equation 1.2, which satisfies $\epsilon-$DP if the following inequality holds.*

$$\sup_{\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{X}^{n-1}: \delta(\mathbf{x},\mathbf{y})=1} \ \sup_{B \in \beta_\Theta} \frac{\xi^{\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x})}{\xi^{\boldsymbol{\alpha}(\mathbf{y})}(B \mid \mathbf{y})} \leq e^{\epsilon}, \tag{1.3}$$

where $\xi^{\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x}) = \int_{\theta \in B} \xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) d\theta$.

Definition 1.3 limits the change in the pseudo posterior distribution over all sets, $B \in \beta_\Theta$ (i.e. β_Θ is the $\sigma-$algebra of measurable sets on Θ), from the inclusion of a single record. Although the pseudo posterior distribution mass assigned to B depends on $\mathbf{x}$, the ϵ guarantee is defined as the supremum over all $\mathbf{x} \in \mathcal{X}^n$ and for all $\mathbf{y} \in \mathcal{X}^{n-1}$ which differ by one record (i.e. $\delta(\mathbf{x}, \mathbf{y}) = 1$).

The α_i's may not be released without leaking information because they are based on the confidential private data, x_i. A draw of modeled parameters, however, may be released along with the synthetic data generated from those parameters (with no leakage of information since all that is released is synthetic data).

Let $\Delta_{\boldsymbol{\alpha},\mathbf{x}} = \sup_{\delta(\mathbf{x},\mathbf{y})=1} \sup_{\theta \in \Theta} |f_\theta^{\alpha}(\mathbf{x}) - f_\theta^{\alpha}(\mathbf{y})|$ be the Lipschitz bound computed, locally, on database $\mathbf{x}$ (over all databases, $\mathbf{y}$, at a LOO distance from $\mathbf{x}$). The pseudo posterior mechanism *indirectly* sets the local DP guarantee, $\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, through the computation of the likelihood weights, $\boldsymbol{\alpha}$.

(author?) (12) show that the local $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ contracts onto the global $\Delta_{\boldsymbol{\alpha}}$, asymptotically in sample size, which in turn drives the contraction of $\epsilon_{\mathbf{x}}$ onto ϵ. For a sample size n sufficiently large, $\epsilon_{\mathbf{x}} = \epsilon$. More formally, the authors demonstrate that the local Lipschitz satisfies a relaxed form of DP that they label "asymptotic DP (aDP)".

We may imagine the generation of multiple collections of databases, $\{\boldsymbol{x}_{n,r}\}_{r=1}^{R}$, that produce the associated collection of Lipschitz bounds, $(\Delta_{\boldsymbol{\alpha},\boldsymbol{x}_{n,r}})_{r=1}^{R}$. The aDP result guarantees that for an n sufficiently large, the local Lipschitz bounds for that collection of databases contract onto the global Lipschitz bound. In practice, **(author?)** (12) show, using a Monte Carlo simulation study for sample sizes of a few hundred, the variability $V(\Delta_{\boldsymbol{\alpha},\boldsymbol{x}_n})$ quickly drops for a specific choice of weight mapping $m()$. We conduct a Monte Carlo simulation study that generates multiple databases to illustrate the asymptotic convergence of a local privacy guarantee to a global privacy guarantee in Appendix D.

More details and Algorithm 2 for implementation are included in Section A.1.

1.4. **Count data privacy alternative for data synthesis.** We formulate a re-weighting adjustment as our main methodological contribution in Section 2 that will adjust the unit indexed weights of the pseudo posterior to improve the utility of the resulting synthetic data produced under the *same* privacy guarantee as achieved before adjustment of the weights. We refer to this improvement in utility for the same privacy guarantee as increasing the "efficiency" of the underlying privatizing mechanism.

(author?) (12) provide a theoretical foundation for the pseudo posterior mechanism, and readers might be left with the question of how to set the $\boldsymbol{\alpha}$ weights in practice. While **(author?)** (12) set each $\alpha_i \in [0, 1]$ to be inversely proportional to the maximum absolute value of the log-likelihood for the record over all θ, there are possibly other ways to measure the disclosure risk r_i for each record other than through the absolute value of the log-likelihood. Next, we introduce an alternative method for measuring risk used to set $\boldsymbol{\alpha}$. There are also multiple options for choosing the decreasing function $m(r_i)$. We show in the sequel that our re-weighting adjustment improves the efficiency under any method for measuring risk. This approach starts with a candidate mapping $m()$ and iterates towards a more optimal mapping $m^*()$ of risks to weights.

The use of the pseudo-posterior for privacy first appeared in the preprint **(author?)** (4), which focused on measures of risk based on categorical data and notions of uniqueness and isolation. (The final published version appears in **(author?)** (5)). This ad-hoc measure of risk was extended to continuous data based on notions of a radius in **(author?)** (7). The authors introduce an alternative method for computing weights $\boldsymbol{\alpha}$ that are estimated as probabilities of identification disclosure, and each $\alpha_i \in [0, 1]$, based on the assumption that a putative intruder guesses randomly from a collection of records whose values are close to or within some set radius of the record being identified.

To compute a weight for each record, $i \in (1, \ldots, n)$, we first calculate its estimated probability of identification disclosure. We assume that an intruder knows the data value of the record she seeks and that she will randomly choose among records that are close to that value. More formally, we cast a ball, $B(y_i, r)$, around the true value of y_i for record i with a radius r. The radius, r, is a policy hyperparameter set by the agency who owns the confidential data. We count the number of records whose values fall *outside* of the radius around the target, and take the ratio of this count over the total number of records, a proportion that we label the risk probability of identification. A target record where the values for most other records lie outside the radius is viewed as isolated because the target

record value is sparsely covered by the values of other records, and therefore at a higher risk of identification disclosure. We then formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, that are inversely proportional to the by-record risk probabilities.

Even though the weights under this scheme are computed based on assumptions about the intruder behavior, we are still able to compute its ($\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$) and invoke a local DP guarantee. The aDP guarantee of **(author?)** (12) requires some conditions that regulate $\boldsymbol{\alpha}$. In particular, the proportion of records receiving reduced weights must be a diminishing proportion. Thus by finding an optimal mapping $m()$, the re-weighting adjustment allows us to attach an aDP property to a broader class of risk-based weights outside of those used for the original pseudo posterior mechanism reviewed in Section 1.3.

More details and Algorithm 3 for implementation are included in Section A.2.

1.5. **Contribution of this paper.** The connection between the pseudo-posterior distribution and a formally private mechanism (asymptotic DP) was first established in the work of **(author?)** (12) building off other published work on the regular posterior mechanism and its connection to DP. This base method was later applied to synthesizing survey data where the contribution was the privatization of both the survey weights and responses (6). The aDP approach has also been combined with an approximate posterior sampling approach leveraging Stochastic Gradient Descent (SGD) to scale up to fine-tuning a pre-trained large language model (1).

Parts of the original **(author?)** (12) implementation may not be satisfactory. For example, the local DP guarantee for a given data set has an asymptotic global property that privacy practitioners may not be comfortable with. This can be strengthened to a global, non-asymptotic DP property by censoring the log-likelihood (e.g bounding the utility function) as demonstrated in **(author?)** (8). However, generating the values from the resulting distribution is more difficult and utility is somewhat reduced (as expected for a stronger privacy guarantee).

In this article, we focus on another challenge with the base implementation: practical aspects of implementing alternative vector-weighted synthesizers, where each alternative synthesizer uses a *different* approach for computing the weights. The main contribution of this paper is to define a recursive "re-weighting" strategy that inputs the vector of privacy weights, α_i's, formulated under any reasonable scheme that defines weights proportionally to the disclosure risks of the data records and subsequently adjusts those weights to achieve a maximally efficient weighting scheme under any level of asymptotic DP privacy guarantee. We use the word "efficient" to denote the minimum distortion of the underlying distribution of the confidential data represented in the released synthetic data for a given privacy guarantee. In other words, we may start with a suboptimal choice of risk r_i mapping of risks to weights $m()$ and iterate to achieve a nearly maximum utility for a fixed privacy guarantee.

We propose a new re-weighting strategy that starts with computation of the maximum of the absolute value of log-pseudo likelihood values over the parameters sampled from the pseudo posterior synthesizer, $\Delta_{\boldsymbol{\alpha},x_i}$, for each data record *after* computing the weights, α_i, and re-estimating the synthesizer under the $\boldsymbol{\alpha}-$weighted pseudo posterior model. The privacy guarantee is driven by the maximum over the data records, $x_i, i \in (1, \ldots, n)$, of the $\Delta_{\boldsymbol{\alpha},x_i}$. So any record, i' with a $\Delta_{\boldsymbol{\alpha},x_{i'}} < \max_{i\in(1,\ldots,n)} \Delta_{\boldsymbol{\alpha},x_i} = \Delta_{\boldsymbol{\alpha},\mathbf{x}}$, is overly downweighted since it does not determine the privacy protection for the overall database. We scale up or increase these weight values, $\alpha_{i'}$, for these overly downweighted data records in a linear

re-weighting step that achieves the same formal privacy guarantee as under the original weights, regardless of the weighting scheme used. The re-weighting strategy improves the utility of the vector-weighted synthesizer while maintaining an equivalent privacy budget. The increased weights, in turn, reduce the distortion encoded for privacy into the released synthetic data, which improves its utility for the user. Lastly, we note that fewer records receiving lower weights provides a tighter connection to the aDP properties in (12). In particular, a diminishing proportion of records receiving down-weighting is needed to ensure convergence. Thus the re-weighting procedure both improves utility and provides stronger justification for the aDP property.

The remainder of the article is organized as follows. We introduce the new re-weighting strategy in Section 2 with an algorithm and a simulation study of highly skewed count data applied to both synthesizers. A Monte Carlo simulation study to demonstrate contraction of local Lipschitz values onto a global Lipschitz is included in Appendix D. We apply our methods to the highly skewed salary variable from a sample of the Survey of Doctorate Recipients in Section 3. Section 4 ends this article with a few concluding remarks.

2. Re-weighting to Maximize Utility for *Any* Vector-weighted Synthesizer

2.1. **Motivation and the proposed method.** To motivate our re-weighting strategy we simulate data from a mixture of two negative binomial distributions, $\text{NB}(\mu = 100, \phi = 5)$ and $\text{NB}(\mu = 100, \phi = 20)$, where ϕ denotes an over-dispersion parameter under which the variance is allowed to be larger than the mean (with mixture weights of $\pi = 0.2$ and $(1-\pi) = 0.8$), which produces data with a highly skewed distribution. All model estimations are performed in Stan (14).

We label the vector-weighted synthesizer in Section 1.3 as LW, which stands for Lipchitz-weighted, and that in Section 1.4 as CW, which stands for Count-weighted. We include the unweighted synthesizer, labeled as "Unweighted", which is a negative binomial synthesizer for the mixture of negative binomial-simulated data.

Figure 1 presents a violin plot of the by-record Lipschitz bounds of each of the two vector-weighted synthesizers, LW and CW, with that of the unweighted synthesizer as a comparison. A violin plot is a density plot rotated 90 degrees and reflected across a vertical axis to emphasize the distribution. We know that the $(\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}})-$aDP privacy guarantee is controlled by the maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$. We see in both violin plots for LW and CW that a substantial mass in the "wings" of the unit level distribution of Lipschitz bounds is below the maximum. The implication is that we are downweighting those units in the wings more than is necesssary to the achieve the $\epsilon_{\mathbf{x}}$ guarantee. We describe this implementation of the weighting schemes as inefficient in that the utility of the resulting synthetic data can be improved for the *same* privacy guarantee if we increase the unit weights for those units in the wings, which would compress this distribution of unit level Lipschitz bounds to more strongly concentrate near the maximum. As long as the maximum of the by-record Lipschitz bounds remains unchanged, we may increase the by-record Lipschitz bounds for other records whose bounds lie below the maximum value to be closer to the overall maximum value without any loss of privacy since the aDP guarantee is based on the maximum Lipschitz bound among the records.

Moreover, in Figure 1 we observe that the weights are more concentrated near the maximum for the LW synthesizer than for the CW. So, the LW is more efficient than the CW

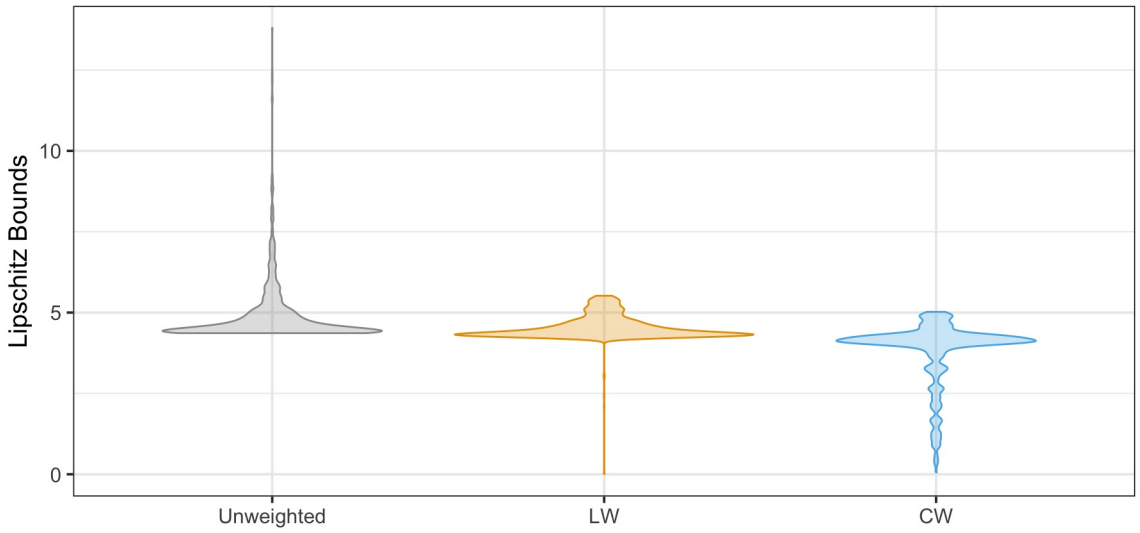


FIGURE 1. Distributions of record-level Lipschitz bounds of the unweighted synthesizer, the LW and CW vector-weighted synthesizers. Both weighted synthesizers create maximum Lipschitz bounds that are significantly lower than the Unweighted. Between them, the LW shows better control of the distribution of Lipschitz bounds with the majority of the by-record Lipschitz bounds being closer to the maximum.

in that the reduced degree of downweighing for most records under LW than CW means that the closely-held data distribution is better preserved since record likelihood contributions are less downweighted under LW.

Figure 2 introduces a second measure of the relative efficiency of a weighting scheme to be used in conjunction with the violin plots of Figure 1. It presents scatter plots of by-record Lipschitz bounds $\Delta_{\boldsymbol{\alpha},x_i}$ (on the y-axis) against by-record weights, α_i, (on the x-axis) for LW and CW. In each case, the red dashed line indicates the maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$. The appearance of a residual relationship between the weights and the Lipschitz bound (blue curves) suggests that an additional refinement to the weights is possible. A lack of residual relationship (close to horizontal line) suggests an optimal weighting. So, we may assess relative efficiencies of weighting schemes by examining the degree of concentration of the mass of unit level Lipschitz values around the maximum and by the degree of deviation of unit Lipschitz values away from the horizontal around the maximum value.

The slopes of the LW and CW Lipschitz lines are relatively similar, but the distributions of the Lipschitz values are more concentrated for LW, indicating that it is an overall more efficient weighting scheme.

Our re-weighting strategy constructs re-weighted weights $\boldsymbol{\alpha}^w = (\alpha_1^w, \cdots, \alpha_n^w)$ by:

$$\alpha_i^w = k \times \alpha_i \times \frac{\Delta_{\boldsymbol{\alpha},\mathbf{x}}}{\Delta_{\boldsymbol{\alpha},x_i}}, \tag{2.1}$$

where $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ is the maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},x_i}$ is the Lipschitz bound for record i, α_i is the weight used in the pseudo posterior synthesizers before the re-weighting step,

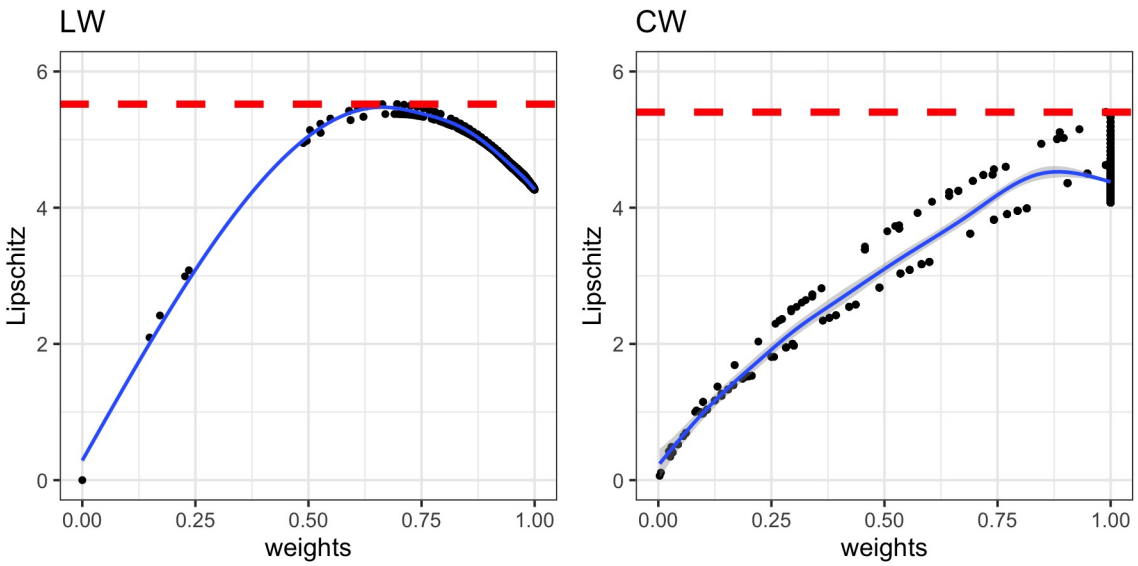


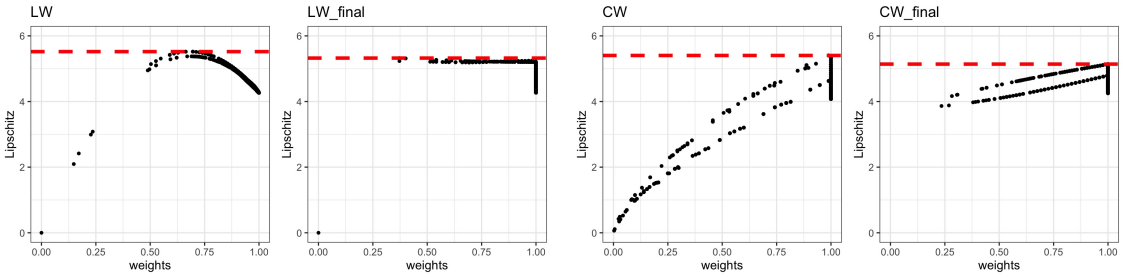
FIGURE 2. Lipschitz Bounds vs Weights, LW and CW. The LW creates the majority of records expressing high Lipschitz bounds due to high weight values. The CW produces many records with low Lipschitz bounds.

and a constant, $k < 1$, is used to ensure that the final maximum Lipschitz bound remains equivalent before and after this re-weighting step. Both $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ and $\Delta_{\boldsymbol{\alpha},x_i}$ are computed from the $\boldsymbol{\alpha}-$weighted pseudo posterior before re-weighting. The implementation for the new re-weighting step is outlined in Algorithm 1. This simple adjustment corrects for suboptimal choice of $m()$ (e.g., the linear transformations used in Algorithms 2 and 3 in Appendix A.1) as well as choice of less efficient risk measures (e.g., LW vs. CW). Under both settings, Algorithm 1 produces an increase in utility for essentially the same privacy-bound.

Algorithm 1: Re-weighting step to obtain $\boldsymbol{\alpha}^w$

1. Use the calculated $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ from the unweighted synthesizer. Use the overall Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, and the by record Lipschitz bounds, $\{\Delta_{\boldsymbol{\alpha},x_i}, i = 1, \cdots, n\}$, computed from the $\boldsymbol{\alpha}-$weighted pseudo posterior synthesizer. Construct a constant $k < 1$ to compute $\boldsymbol{\alpha}^w$, where each $\alpha_i^w = k \times \alpha_i \times \frac{\Delta_{\boldsymbol{\alpha},\mathbf{x}}}{\Delta_{\boldsymbol{\alpha},\mathbf{x}_i}} \in [0, 1]$.;
2. Run step 5 to 7 in Algorithm 2 / Algorithm 3, again, to re-estimate the synthesizers under an $\boldsymbol{\alpha}^w-$weighted pseudo posterior to obtain $\Delta_{\boldsymbol{\alpha}^w,\mathbf{x}}$.;
3. Stop if $|\Delta_{\boldsymbol{\alpha},\mathbf{x}} - \Delta_{\boldsymbol{\alpha}^w,\mathbf{x}}| < \tau$ for some tolerance $\tau > 0$. If $\Delta_{\boldsymbol{\alpha},\mathbf{x}} < \Delta_{\boldsymbol{\alpha}^w,\mathbf{x}}$, reduce k. Otherwise increase k.;
4. Repeat steps 1 to 3.

2.2. **Application to simulated highly skewed data.** We demonstrate our re-weighting strategy under the highly skewed negative binomial mixture data. Using $k = 0.95$ produces



(A) LW (B) CW

FIGURE 3. Lipschitz Bounds vs Weights, before and after re-weighting for LW (panel A) and CW (panel B). The re-weighting strategy produces a nearly horizontal Lipschitz-weight relationship. The impact is more significant for LW.

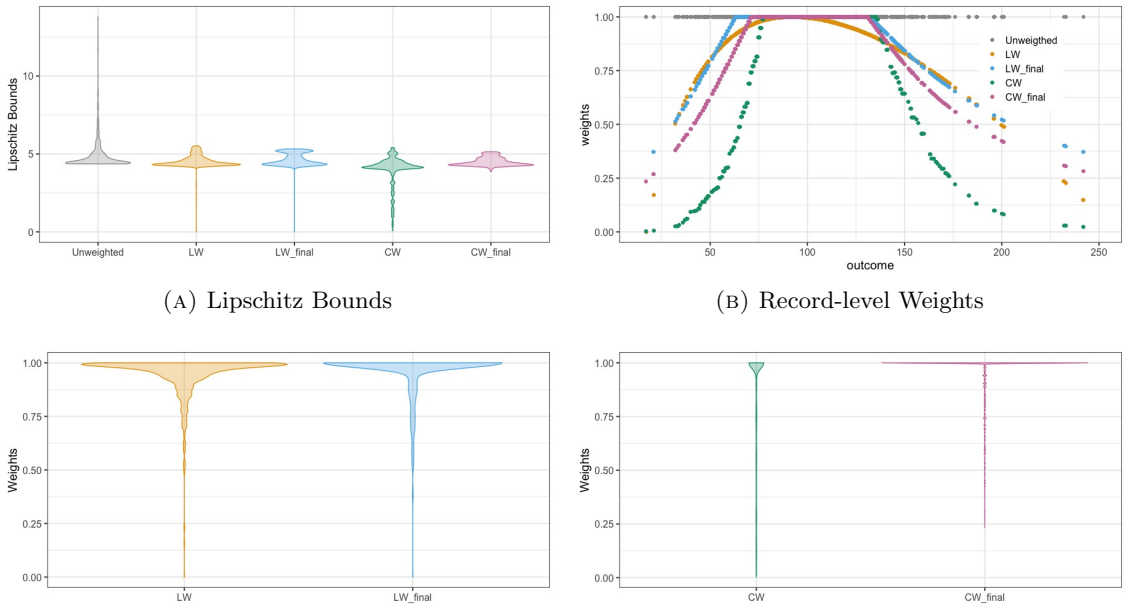
an equivalent overall Lipschitz bound. The re-weighted synthetic data results are labeled "LW_final" and "CW_final", for LW and CW respectively.

Figure 3a and Figure 3b show the before vs after re-weighting scatter plots of Lipschitz bounds and weights. As is evident in Figure 3a, the curve showing the Lipschitz-weight relationship becomes nearly horizontal at the maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ as we move from LW to LW_final, which indicates maximum efficiency. The curve in Figure 3b becomes much shallower from CW to CW_final, indicating much improved efficiency.

Figure 4a confirms that our re-weighting strategy increases the weighting efficiency of LW and CW in that the by-record Lipschitz bounds are increased while maintaining an equivalent maximum Lipschitz bound. At the record level, Figure 4b illustrates that every record has received a higher weight from LW to LW_final, and from CW to CW_final. We receive further confirmation of the improved efficiency of implementing the re-weighting step from the weight plots in Figure 4c and Figure 4d that show weights increase after re-weighting.

Turning to the utility performances, Figure 5a and Figure 5b show notable improvement in utility of CW after re-weighting for all estimates of generated synthetic data. The deterioration in the preservation of the real (not synthetic) data distribution tails due to overly downweighting records in the tails *before* re-weighting is greatly mitigated by the re-weighting strategy. We observe that all of the extreme quantiles, the mean, and the median estimates are much more accurate. The improvement of utility of LW is less impressive because the relative improvement in the efficiency of the weighting scheme is relatively smaller. However, we can certainly see improvement in estimating the mean, median, and 90th quantile. For example, compared to a 95% confidence interval of median in the data [96.0, 101.0], the CW_final achieved [96.1, 100.1] improved from CW's [98.7, 102.4], and LW_final achieved [96.0 100.0] improved from LW's [95.6, 99.7]. We include a table of comparisons of all estimands in Appendix B.

Moreover, utility of the mean parameter μ estimation in Figure 5c improves after re-weighting for LW and CW, with a bigger improvement for CW. In Figure 5d, we turn to a violin plot for the data distribution compared to the synthesizer distributions. We see that the re-weighting strategy improves the preservation of the confidential data distribution tail in LW_final and CW_final, compared to LW and CW, respectively. This reduced downweighting



(A) Lipschitz Bounds

(B) Record-level Weights

(C) LW Weights

(D) CW Weights

FIGURE 4. Before vs after re-weighting for the negative binomial mixture: Lipschitz Bounds (panel A) and Weights (panels B to D). After re-weighting, the by-record Lipschitz bounds are closer to the maximum, whereas the record-level weights are closer to one, indicating the effectiveness of the re-weighting strategy to increase the weighting efficiency of both LW and CW.

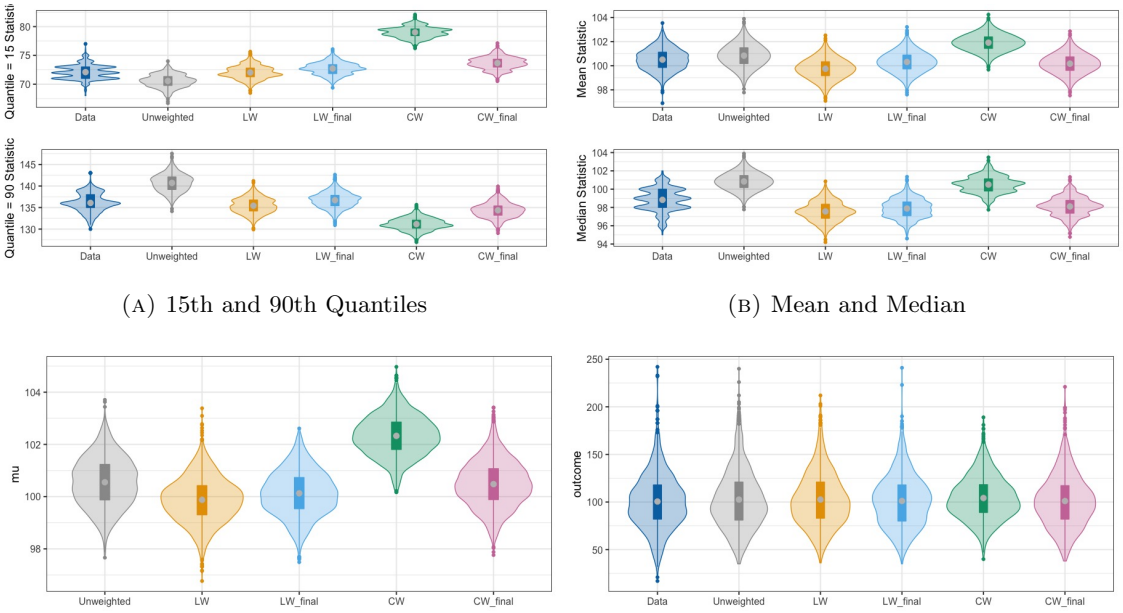
of the distribution tails is a feature of the re-weighting strategy for any vector-weighted scheme applied to highly skewed data.

3. Application to The Survey of Doctorate Recipients

The Survey of Doctorate Recipients (SDR) provides demographic, education, and career history information from individuals with a U.S. research doctoral degree in a science, engineering, or health (SEH) field. The SDR is sponsored by the National Center for Science and Engineering Statistics and by the National Institutes of Health. In this section, we demonstrate our re-weighting strategy on a sample of 1000 observations of the SDR focused on the highly skewed salary variable. The sample comes from the 2017 Survey of Doctorate Recipients public use file (`https://ncses.nsf.gov/explore-data/microdata`). The highly skewed salary variable has a mean of \$107,609, a median of a \$95,000, a range of [\$0, \$509,000], and a standard deviation of \$69,718. We use a negative binomial unweighted synthesizer for this highly skewed variable salary.

3.1. Before re-weighting.

Before re-weighting, the results of LW and CW on the real skewed data sample tell a similar story as on simulated skewed data. The results are included in the Appendix C for brevity and we summarize the findings here: i) LW has the highest



(A) 15th and 90th Quantiles

(B) Mean and Median

(C) Posterior Density of μ

(D) Synthetic Data

FIGURE 5. Utility Measures Before vs after re-weighting for the negative binomial mixture: tail quantiles (panel A) and central measures (panel B) for synthetic data, parameter estimates (panel C), and overall distribution for synthetic data (panel D). The re-weight strategy improves the utility of several quantities of interest for both LW and CW, with a more significant improvement for CW.

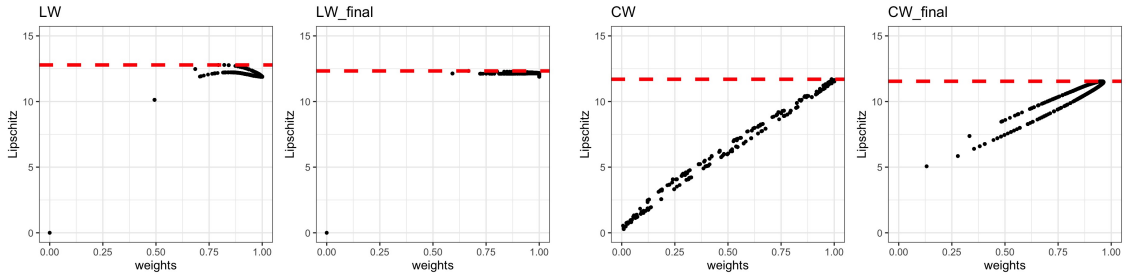
utility among the three, though its relatively heavy downweighting of records in the tails of the confidential data distribution under highly skewed data results in reduced utility compared to that of less skewed data; ii) CW's utility performance on the real skewed data is worse than that on the simulated skewed data – it has assigned low weights to many more records, resulting in low weighting efficiency and therefore low utility.

LW's weighting efficiency is close to optimal for most data records such that re-weighting might not improve much. However, CW's weighting efficiency is expected to see large improvement after re-weighting, which in turn, will improve its utility performance.

3.2. **After re-weighting.** We apply the re-weighting strategy to LW and CW to maximize their utility performances. We set $k = 0.95$ to maintain an equivalent overall Lipschitz bound. The results are labeled as "LW_final" and "CW_final" respectively.

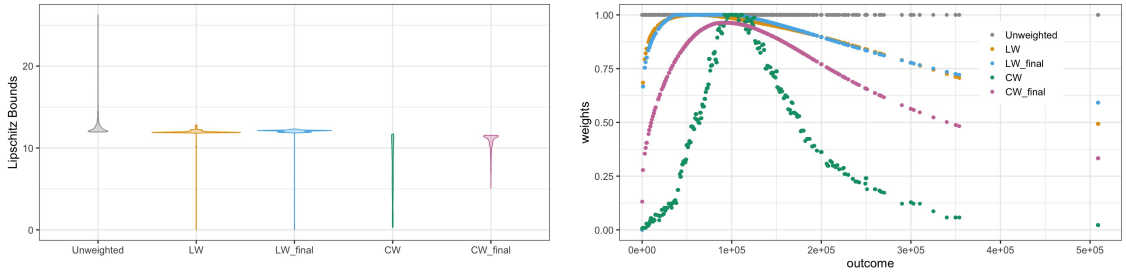
Figure 6a shows that the re-weighting strategy has pushed the Lipschitz-to-weight association to almost horizontal at the maximum Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, for LW, indicating maximum efficiency. Figure 6b shows that the re-weighting strategy has also produced a Lipschitz-to-weight association that is less vertical for CW. Therefore, we expect to see minor utility improvement of LW_final, and a sizable utility improvement of CW_final.

Examining the distributions for the weights under both of LW and CW in Figure 7 shows how much weighting efficiency LW_final and CW_final have gained after the re-weighting

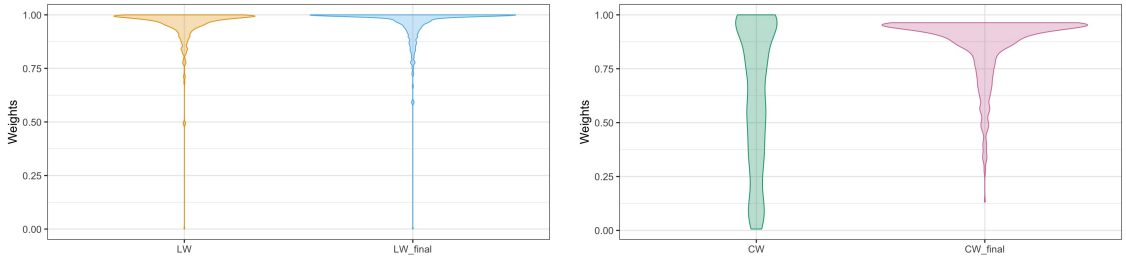


(A) LW (B) CW

FIGURE 6. Lipschitz Bounds vs Weights, before and after re-weighting for LW (panel A) and CW (panel B). The re-weighting strategy produces a nearly horizontal Lipschitz-weight relationship. The impact is more significant for the LW.



(A) Lipschitz Bounds (B) Record-level Weights



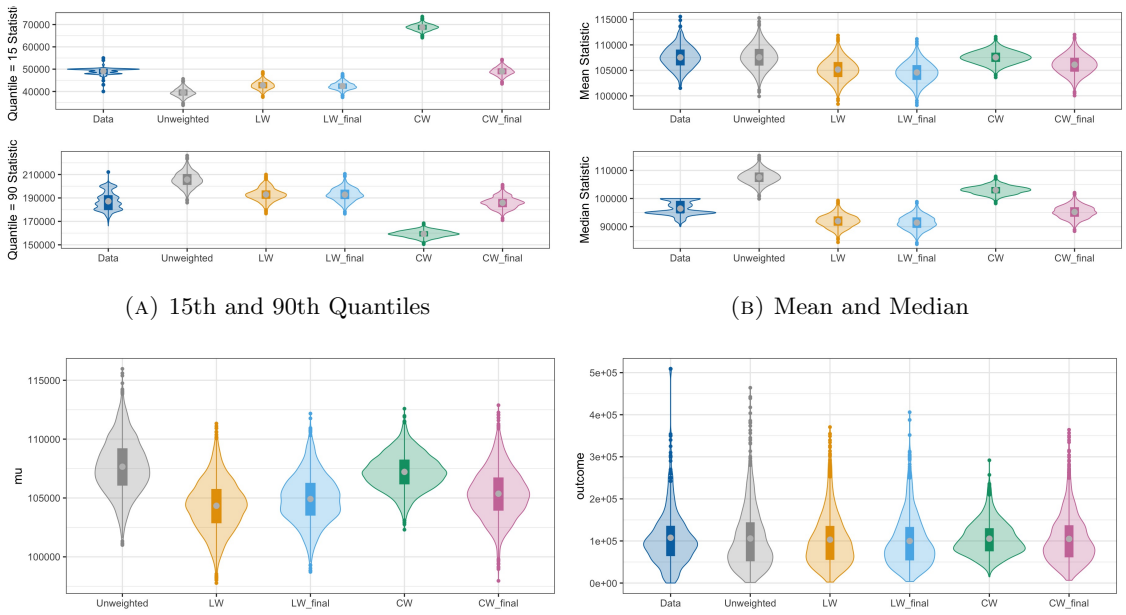
(C) LW Weights (D) CW Weights

FIGURE 7. Before vs after re-weighting for salary: Lipschitz Bounds (panel A) and Weights (panels B to D). After re-weighting, the by-record Lipschitz bounds are closer to the maximum, whereas the record-level weights are closer to one, indicating the effectiveness of the re-weighting strategy to increase the weighting efficiency of both LW and CW.

strategy. Focusing on the walk between CW to CW_final in Figure 7d, the distribution of weights is highly diffuse with large mass assigned to low values before re-weighting. After

re-weighting, by contrast, many more records receive higher weights. Even though the by-record weights ($\boldsymbol{\alpha}$) have increased after re-weighting, Figure 7a shows that the re-weighting strategy has maintained an equivalent maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$.

Finally, the utility results in Figure 8 demonstrate the utility maximizing feature of our proposed re-weighting strategy on the confidential data sample. Whether it is the preservation of statistics of the confidential data distribution, shown in Figure 8a and Figure 8b, the relative accuracy of parameter estimates in Figure 8c, or similarity of the synthetic data density to that of the confidential data in Figure 8d, CW_final has produced much higher utility than CW across the board, a result that we expect to see given its improved weighting efficiency previously discussed. We make particular mention that the comparisons of the confidential and synthetic data distributions in Figure 8d show that re-weighting reduces or mitigates the shrinking of the tail of the confidential data in the resulting synthetic data. Overall, there is a minor utility improvement from LW to LW_final, another result we expect to see given its minor improvement of weighting efficiency. Nevertheless, our proposed re-weighting strategy maximizes the utility of any vector-weighted scheme, while maintaining an equivalent maximum Lipschitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$.



(A) 15th and 90th Quantiles

(B) Mean and Median

(C) Posterior Density of μ

(D) Synthetic Data

FIGURE 8. Utility Measures Before vs after re-weighting for salary: tail quantiles (panel A) and central measures (panel B) for synthetic data, parameter estimates (panel C), and overall distribution for synthetic data (panel D). The re-weight strategy improves the utility for both LW and CW.

4. Concluding Remarks

In this article, we address the issue of how to choose an optimal weighting strategy for the pseudo posterior synthesizer, which involves both a risk measure and a mapping of

risks to weights. We introduce a new re-weighting strategy that improves utility of *any* vector-weighted scheme in the difficult case of a highly-skewed data distribution, while maintaining an equivalent privacy budget. We demonstrate through the application to two different risk measures (one more efficient and one less efficient) and a naive linear mapping that this strategy improves weighting efficiency by increasing by-record weights to compress the distribution of by-record Lipschitz bounds. Improved weighting efficiency substantially mitigates the tendency for vector-weighted schemes to overly downweight the tails, especially for risk measures that are less efficient (e.g., not directly based on the likelihood contributions). Weighting efficiency also leads to fewer records receiving reduced weights, yielding improved convergence and providing a stronger justification for the asymptotic privacy guarantee of (12). Existing implementations such as those used for synthesizing survey weighted data (6) and privately training large language models (1) can immediately benefit from this improved weighting approach.

References

[1] Robert Chew, Matthew R. Williams, Elan A. Segarra, Alexander J. Preiss, Amanda Konet, and Terrance D. Savitsky. Bayesian pseudo posterior mechanism for differentially private machine learning. *arXiv preprint*, 2025. arXiv:2503.21528, https://doi.org/10.48550/arXiv.2503.21528.

[2] C. Dimitrakakis, B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *Journal of Machine Learning Research*, 18(1):343–381, 2017. URL: http://jmlr.org/papers/v18/15-257.html.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. https://doi.org/10.1007/11681878_14.

[4] Jingchen Hu and Terrance D. Savitsky. Bayesian data synthesis and disclosure risk quantification: An application to the consumer expenditure surveys. *arXiv preprint*, 2021. arXiv:1809.10074, https://doi.org/10.48550/arXiv.1809.10074.

[5] Jingchen Hu and Terrance D. Savitsky. Bayesian data synthesis and disclosure risk quantification: An application to the consumer expenditure surveys. *TRANSACTIONS ON DATA PRIVACY*, 16:83–121, 2023. URL: https://www.tdp.cat/issues21/tdp.a437a21.pdf.

[6] Jingchen Hu, Terrance D Savitsky, and Matthew R Williams. Private tabular survey data products through synthetic microdata generation. *Journal of Survey Statistics and Methodology*, 10(3):720–752, 2022. https://doi.org/10.1093/jssam/smac001.

[7] Jingchen Hu, Terrance D Savitsky, and Matthew R Williams. Risk-Efficient Bayesian Data Synthesis for Privacy Protection. *Journal of Survey Statistics and Methodology*, 10:1370–1399, 2022. https://doi.org/10.1093/jssam/smab013.

[8] Jingchen Hu, Matthew R Williams, and Terrance D Savitsky. Mechanisms for global differential privacy under bayesian data synthesis. *Statistica Sinica*, 35:563–584, 2025. https://doi.org/10.5705/ss.202022.0162.

[9] R. J. A. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426, 1993. URL: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

[10] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007. https://doi.org/10.1109/FOCS.2007.66.

[11] D. B. Rubin. Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9:461–468, 1993. URL: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

[12] Terrance D. Savitsky, Matthew R. Williams, and Jingchen Hu. Bayesian pseudo posterior mechanism under asymptotic differential privacy. *Journal of Machine Learning Research*, 23:1–37, 2022. URL: https://www.jmlr.org/papers/volume23/21-0936/21-0936.pdf.

[13] J. Snoke and A. Slavkovic. pMSE mechanism: Differentially private synthetic data with maximal distributional similarity. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, volume 11126 of *Lecture Notes in Computer Science*, pages 138–159. Springer, 2018. https://doi.org/10.1007/978-3-319-99771-1_10.

[14] Stan Development Team. RStan: the R interface to Stan, 2025. R package version 2.32.7. URL: https://mc-stan.org/.

[15] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2493–2502, Lille, France, 07–09 Jul 2015. PMLR. URL: http://proceedings.mlr.press/v37/wangg15.html.

[16] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105:375–389, 2010. https://doi.org/10.1198/jasa.2009.tm08651.

Appendix A. Vector-weighted Synthesizer Algorithms

The first synthesizer, labeled as likelihood-weighted (LW), is the mechanism of (12) described above. The second synthesizer, labeled as count-weighted (CW), sets each by-record weight, $\alpha_i \in [0,1]$, such that $\alpha_i = m(IR_i) \propto 1/IR_i$, where IR_i denotes the disclosure risk probability ($\in [0,1]$) of record i. The IR_i is a measure of a record's isolation from other records and is constructed by counting the number of records whose values are outside a radius around the true value for the target record divided by the total number of records (7). A record whose true value is not well-covered by the values of other records is relatively more isolated and, therefore, at higher disclosure risk. The radius is a measure of closeness that is tuned by the owner of the confidential data.

In both synthesizers, the record-indexed vector weights $\boldsymbol{\alpha} = (\alpha_1 \in [0,1], \cdots, \alpha_n \in [0,1])$ are used to exponentiate the likelihood contributions where the weights are designed to target high-risk records by downweighting their likelihood contributions. These two measures of risk, LW and CW, are related in that the notion of record isolation underlies both. The value of the response variable for an isolated target record is near to or within a close radius to the values of many other records. As earlier mentioned, such isolated records generally appear in the tails of the distribution where there is little distribution mass. As a result, downweighting the likelihood contribution of isolated records tends to preserve the high mass regions of the confidential data distribution in the resulting synthetic data generated for release.

We specify the method for formulation of vector weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ for the LW and CW synthesizers in Section 1.3 and Section 1.4, with Algorithm 2 and Algorithm 3, respectively.

Each algorithm starts by computing the weights, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, which are then used to construct the pseudo likelihood and estimate the pseudo posterior. Next, we draw parameters from the estimated pseudo posterior distribution and compute the overall Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ for database, $\mathbf{x}$. The resulting DP guarantee is $(\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}})$ and is "local" to the database, $\mathbf{x}$, and the $\epsilon_{\mathbf{x}}$ is *indirectly* controlled through the weights, $\boldsymbol{\alpha}$. Synthetic data are then generated using the drawn $(\theta_s)_{s=1,\ldots S}$ from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution from step 5 in each algorithm of the corresponding data generating model.

As earlier discussed, the local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, contracts onto the "global" Lipschitz bound, $\Delta_{\boldsymbol{\alpha}}$, over all databases, $\mathbf{x} \in \mathcal{X}^n$ of size n, as n increases such that $\epsilon_{\mathbf{x}}$ contracts onto ϵ at a relatively modest sample size.

A.1. **Generating synthetic data under the LW synthesizer.** We specify the algorithm for generating synthetic data under the LW $\boldsymbol{\alpha}-$weighted pseudo posterior distribution that produces synthetic data for database, $\mathbf{x}$.

To implement the LW vector-weighted synthesizer, we first fit an unweighted synthesizer and obtain the absolute value of the log-pseudo likelihood for each data base record i and each Markov chain Monte Carlo (MCMC) draw s of θ from the unweighted posterior distribution. A Lipschitz bound for each record is computed by taking the maximum of the log-likelihoods over the S posterior draws of θ. We formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, to be inversely proportional to those by-record Lipschitz bounds. See step 1 to step 4 in Algorithm 2.

Algorithm 2 is implemented on the observed database, $\mathbf{x} \in \mathcal{X}^n$, under which we compute the local (specific-to-database $\mathbf{x}$) Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, to achieve a local privacy guarantee, $\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, which is equivalent to an $\epsilon = \epsilon_{\mathbf{x}}-$aDP guarantee where $\epsilon = \epsilon_{\mathbf{x}}$ for an n sufficiently large.

We emphasize that LW *indirectly* achieves the aDP guarantee, $(\epsilon = 2\Delta_{\alpha})$, through the computation of the likelihood weights, $\boldsymbol{\alpha}$. The LW algorithm constructs weights intended to directly minimize the overall Lipschitz bound for the synthetic data by downweighting the likelihood contribution for each record in inverse proportion to the absolute value of its log-likelihood. We recall that the Lipschitz is the maximum over the parameter space and records of this absolute value of the log-likelihood quantity, so our LW weighting scheme will be efficient at targeting those high risk records that most effect the privacy guarantee to produce a relatively moderate distortion of the confidential data distribution expressed in the publicly-released synthetic data.

A.2. **Generating synthetic data under the CW synthesizer.** Next, we present the algorithm of **(author?)** (7) for generating synthetic data under the CW $\boldsymbol{\alpha}-$weighted synthesizer. The weights $\boldsymbol{\alpha}$ are estimated as probabilities of identification disclosure, and each $\alpha_i \in [0, 1]$, based on the assumption that a putative intruder guesses randomly from a collection of records whose values are near to or within some set radius of the record being identified.

Algorithm 2: Steps to implement the LW vector-weighted synthesizer

1. Let $|f_{\theta_s,i}|$ denote the absolute value of the log-likelihood computed from the unweighted synthesizer for database record, $i \in (1, \ldots, n)$ and MCMC draw, $s \in (1, \ldots, S)$ of θ from its unweighted posterior distribution;
2. Compute the $S \times n$ matrix of by-record absolute value of log-likelihoods, $L = \{|f_{\theta_s,i}|\}_{i=1,\ldots,n,\ s=1,\ldots,S}$;
3. Compute the maximum over each $S \times 1$ column of L to produce the $n \times 1$ (database record-indexed) vector, $\mathbf{f} = (f_1, \ldots, f_n)$. We use a linear transformation of each f_i to $\tilde{f}_i \in [0, 1]$ where values of $\tilde{f}_i$ closer to 1 indicates relatively higher identification disclosure risk: $\tilde{f}_i = \frac{f_i - \min_j f_j}{\max_j f_j - \min_j f_j}$;
4. Formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, $\alpha_i = c \times (1 - \tilde{f}_i) + g$, where c and g denote scaling and shift parameters, respectively, of the α_i used to tune the risk-utility trade-off for setting $\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$;
5. Use $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ to construct the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution, $\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \propto \prod_{i=1}^{n} p(x_i \mid \theta)^{\alpha_i} \times \xi(\theta)$. Draw $(\theta_s)_{s=1,\ldots S}$ from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution, where S denotes the number of draws of θ from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution;
6. Compute the $S \times\ n$ matrix of log-pseudo likelihood values, $L^{\boldsymbol{\alpha}} = \left\{|f^{\alpha_i}_{\theta_s,i}|\right\}_{i=1,\ldots,n,\ s=1,\ldots,S}$ where $f^{\alpha_i}_{\theta_s,i} = \log p(x_i \mid \theta_s)^{\alpha_i}$;
7. Compute $\Delta_{\boldsymbol{\alpha},\mathbf{x}} = \max_{s,i}|f^{\alpha_i}_{\theta_s,i}|$, that defines the local DP guarantee, $\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, for database $\mathbf{x}$.

To compute a weight for each record, $i \in (1, \ldots, n)$, we first calculate its estimated probability of identification disclosure. We assume that an intruder knows the data value of the record she seeks and that she will randomly choose among records that are close to that value. More formally, we cast a ball, $B(y_i, r)$, around the true value of y_i for record i with a radius r. The radius, r, is a policy hyperparameter set by the agency who owns the confidential data. We count the number of records whose values fall *outside* of the radius around the target, and take the ratio of this count over the total number of records, a proportion that we label the risk probability of identification. A target record where the values for most other records lie outside the radius are viewed as isolated because the target record value is sparsely covered by the values of other records, and therefore at a higher risk of identification disclosure. We then formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, that are inversely proportional to the by-record risk probabilities. See step 1 to step 4 in Algorithm 3.

Even though the weights under CW are computed based on assumptions about the intruder behavior, we are still able to compute its ($\epsilon_{\mathbf{x}} = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$) and invoke a local DP guarantee. The aDP guarantee of **(author?)** (12) requires some conditions that regulate $\boldsymbol{\alpha}$. In particular, the proportion of records receiving reduced weights must be a diminishing proportion. Thus the re-weighting approach allows us to attach an aDP property to a broader class of risk-based weights outside of the LW.

Algorithm 3: Steps to implement the CW vector-weighted synthesizer.

1. Let M_i denote the set of records in the original data, and $|M_i|$ denote the number of records in the set;
2. Cast a ball, $B(y_i, r)$ with a radius r around the true value of record i, and count the number of records falling outside the radius $\sum_{j \in M_i} \mathbb{I}\left(y_j \notin B(y_i, r)\right)$;
3. Compute the record-level risk probability, IR_i as $IR_i = \sum_{j \in M_i} \mathbb{I}\left(y_j \notin B(y_i, r)\right) / |M_i|$, such that $IR_i \in [0, 1]$;
4. Formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$, $\alpha_i = c \times (1 - IR_i) + g$, where c and g denote scaling and shift parameters, respectively, of the α_i used to tune the risk-utility trade-off;
5. Use $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ to construct the pseudo likelihood from which the pseudo posterior is estimated. Draw $(\theta_s)_{s=1,\ldots S}$ from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution;
6. Compute the $S \times \ n$ matrix of log-pseudo likelihood values, $L^{\boldsymbol{\alpha}} = \left\{ |f_{\theta_s,i}^{\alpha_i}| \right\}_{i=1,\ldots,n,\ s=1,\ldots,S}$;
7. Compute $\Delta_{\boldsymbol{\alpha},\mathbf{x}} = \max_{s,i} |f_{\theta_s,i}^{\alpha_i}|$, that defines the local DP guarantee for database $\mathbf{x}$.

Appendix B. Utility comparison before and after re-weighting in Section 2.2

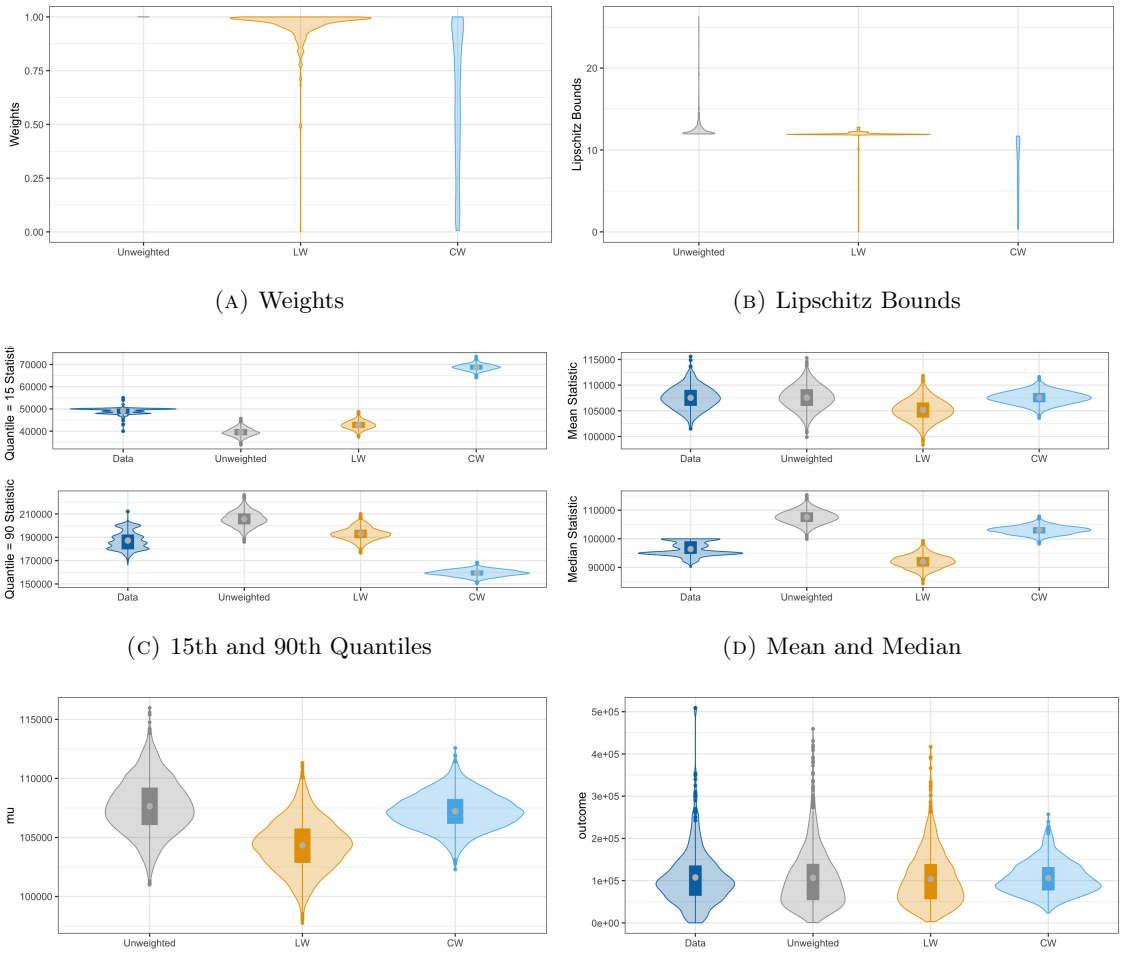
Table 1 presents utility comparison of LW and CW before and after re-weighting in the simulation studies in Section 2.2.

	Data	LW	LW_final	CW	CW_final
15th quantile	[70.0, 75.0]	[70.0, 74.3]	[70.6, 74.8]	[77.2, 80.8]	[71.6, 75.8]
90th quantile	[132.0, 140.0]	[132.1, 139.0]	[133.1, 140.3]	[128.5, 134.0]	131.0, 137.7]
mean	[98.8, 102.3]	[98.0, 101.4]	[98.7, 102.0]	100.6, 103.3]	[98.6, 101.8]
median	[96.0, 101.0]	[95.6, 99.7]	[96.0, 100.0]	[98.7, 102.4]	[96.1, 100.1]

Table 1. Utility comparison before and after re-weighting for the negative binomial mixture: 95% confidence interval.

Appendix C. Plots of LW and CW before re-weighting in Section 3.1

Figure 9 provides results of LW and CW *before* re-weighting for the application in Section 3.1.



(A) Weights (B) Lipschitz Bounds

(C) 15th and 90th Quantiles (D) Mean and Median

(E) Posterior Density of μ (F) Synthetic Data

FIGURE 9. Violin Plots for Salary by Weighting approach: Weight distributions (panel A), Lipschitz Bounds (panel B), tail quantiles (panel C) and central measures (panel D) for synthetic data, parameter estimates (panel E), and overall distribution for synthetic data (panel F).

APPENDIX D. MOVING FROM LOCAL-TO-GLOBAL PRIVACY GUARANTEE

We proceed to implement a Monte Carlo simulation study under each of our less skewed Poisson and more skewed mixture of negative binomials data generating models to walk from the local privacy guarantee for a specific database, $\mathbf{x}$, to a global (asymptotic DP) guarantee over the space of databases, $\forall \mathbf{x} \in \mathcal{X}^n$. We generate $R = 100$ local databases under each generating model, estimate the unweighted and LW weighted synthesizers on each database, $r \in (1, \ldots, (R = 100))$, and compute a local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}_r}$, on each $\mathbf{x}_r$ under each synthesizer. We plot the distributions of the $(\Delta_{\boldsymbol{\alpha},\mathbf{x}_r})_{r=1}^{R=100}$ for each synthesizer and conclude

that we have achieved a global asymptotic DP result if this distribution contracts around a global $\Delta_{\boldsymbol{\alpha}}$. We summarize our Monte Carlo simulation procedure, below:

(1) For $r = 1, \ldots, (R = 100)$:
   - Generate $\mathbf{x}_r \sim \text{Pois}(\mu)$ or $\mathbf{x}_r \sim \pi_1 \text{NB}(\mu_1 = 100, \phi_1 = 5) + \pi_2 \text{NB}(\mu_2 = 100, \phi_2 = 5)$, each of size $n = 1000$.
   - Compute the *local* Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}_r}$, for the unweighted and $\boldsymbol{\alpha}-$re-weighted synthesizers.
   - Construct the distribution of $\Delta_{\boldsymbol{\alpha},\mathbf{x}_r}$ and note the maximum of the distribution and difference between the maximum and minimum values of the distribution of the local Lipschitz bounds.

(2) Assess contraction of the $\max_r \Delta_{\boldsymbol{\alpha},\mathbf{x}_r}$ to a single (global) value and whether the minimum and maximum values collapse together.

Figure 10 presents a violin plot of the local $(\Delta_{\boldsymbol{\alpha},\mathbf{x}_r})_{r=1}^R$ for the $R = 100$ Monte Carlo iterations of the less skewed data generating model for the Unweighted (left) and LW-weighed (right) synthesizers. We readily observe that the LW synthesizer contracts onto the global value, $\epsilon = 2\Delta_{\boldsymbol{\alpha}} = 7$. That this contraction is consistent with a relaxed, aDP guarantee comes from the small distribution mass above $\Delta_{\boldsymbol{\alpha}} = 3.5$, though we see the probability that the Lipschitz bound for any local database that exceeds 3.5 is nearly 0 at $n = 1000$.

Figure 11 presents the associated distributions of a set of estimands over the Monte Carlo iterations for each synthesizer as compared to the confidential data. There is an expected loss of utility, though inference is reasonably well-preserved.

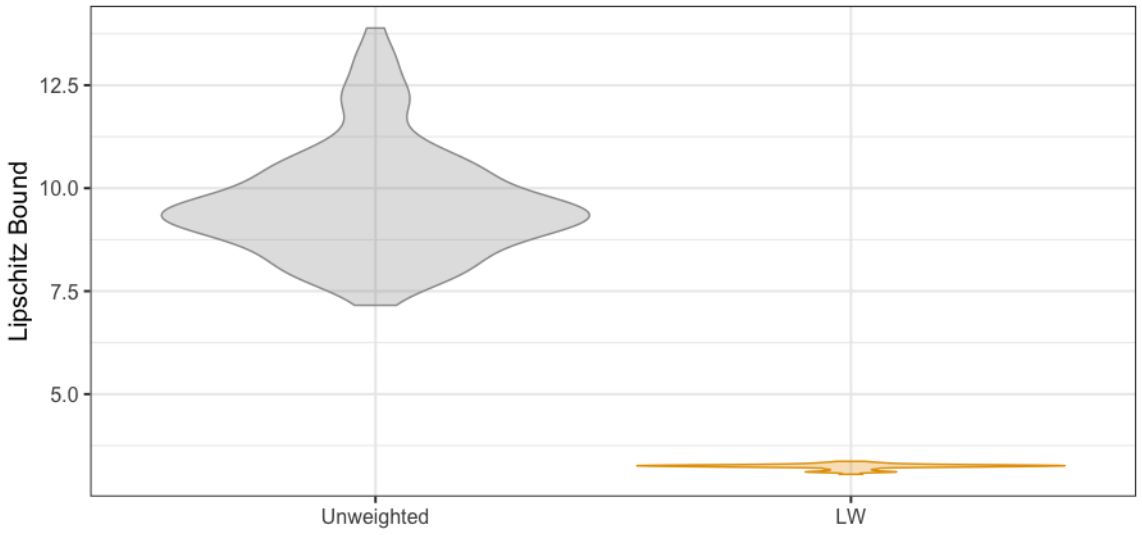
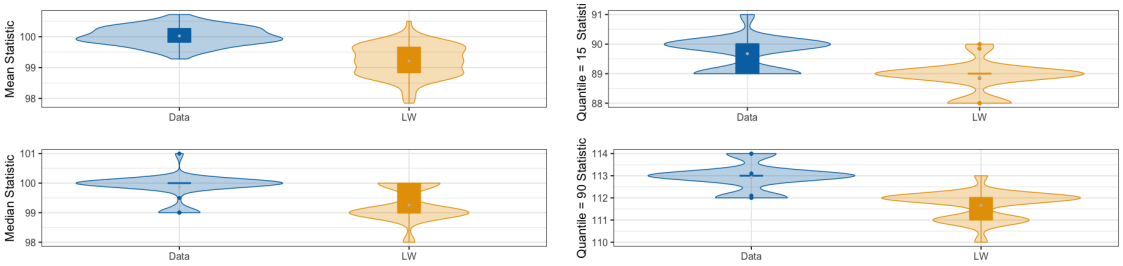


FIGURE 10. Violin plot of Lipschitz bounds over the Monte Carlo iterations under the Poisson generating model.



(A) Mean and Median (B) 15th and 90th Quantiles

FIGURE 11. Violin Plots for the estimands of the confidential and synthetic data distributions over the Monte Carlo iterations for the Poisson generating model. Central measures (panel A) and tail quantiles (panel B).

The following set of figures repeat the earlier set, but here under the highly skewed data generation process from a negative binomial mixture. The conclusions are the same as in the earlier set when we observed substantial contraction around the global Lipschitz (where here $\Delta_{\boldsymbol{\alpha}} = 6$).

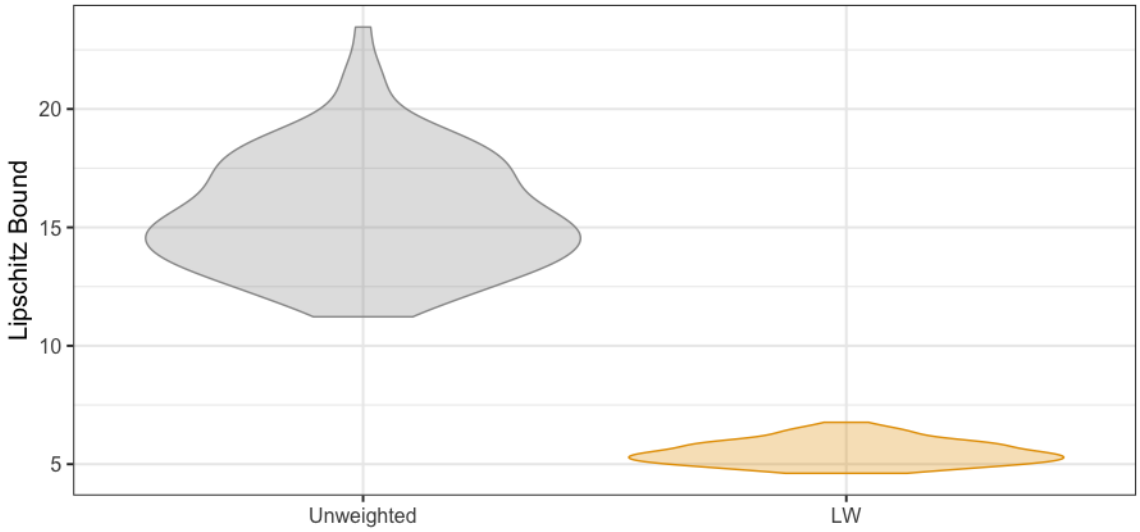
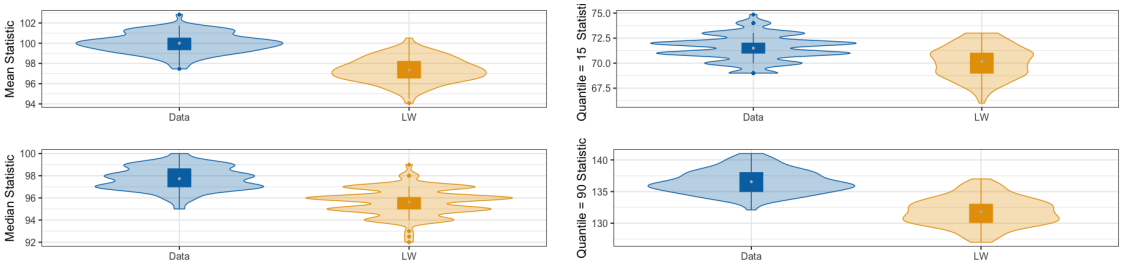


FIGURE 12. Violin plot of Lipschitz bounds over the Monte Carlo iterations under the negative binomial mixture generating model.



(A) Mean and Median (B) 15th and 90th Quantiles

FIGURE 13. Violin Plots for the estimands of the confidential and synthetic data distributions over the Monte Carlo iterations for the mixtures of negative binomials generating model. Central measures (panel A) and tail quantiles (panel B).

We have illustrated the theory of **(author?)** (12) that guarantees an asymptotic DP result by demonstrating a contraction of a collection of Lipschitz bounds for local databases onto a global Lipschitz bound under both of our low and highly skewed data generating scenarios, with our LW synthesizer and re-weighting strategy. The key conclusion is that for an n sufficiently large, a local Lipschitz bound estimated on a specific database, $\mathbf{x}$, becomes arbitrarily close to the global Lipschitz bound.