Privacy Amplification for Synthetic data using Range Restriction

Monika (Jingchen) Hu[1] Matthew R. Williams[2]
Terrance D. Savitsky[3]

[1] Binghamton University (Department of Mathematics and Statistics)
[2] RTI International (Center for Official Statistics)
[3] U.S. Bureau of Labor Statistics (Office of Survey Methods Research)

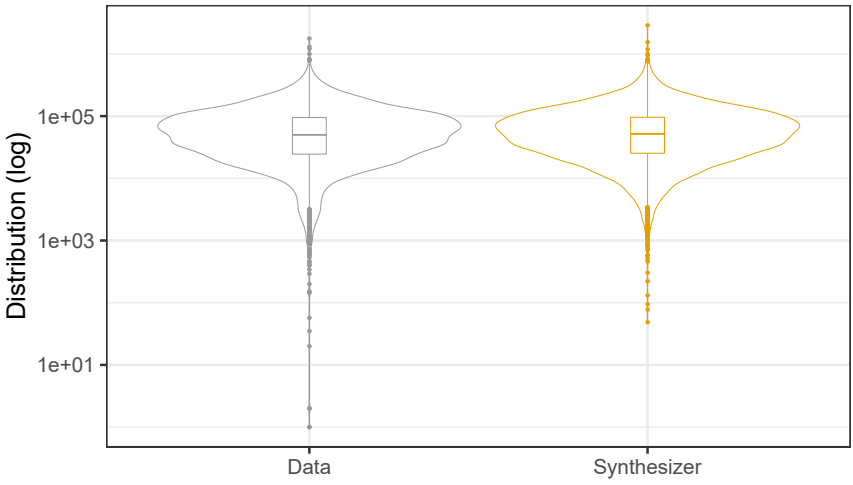
2026 Privacy and Public Policy Conference

Outline

Synthetic data

- Respondent-level data generation and release
- Estimate a model on the confidential microdata
- Simulate synthetic data from posterior predictive distribution
- Constructing tables is one way to measure synthetic data utility
- Unlimited queries; e.g. survey tables, regression analyses etc.
- High utility and low risks (non-formal)

Replicate data $\mathbf{y}^*$ given (|) observed data, $\mathbf{x}$ (Hu, 2019)



Formal Privacy

- Formal Privacy is a property associated with release mechanisms $\mathcal{M}$ (e.g. generating a synthetic data set).
  - has a mathematical bound on the information contribution of individuals
  - allows for privacy accounting (e.g. summing privacy loss/expense across all uses of the data)
  - requires that the mechanism has to be random (stochastic)
- The most popular formal privacy definition (family) is differential privacy (DP), where the bounds are over neighboring data sets with and without individual data

Why Differential Privacy (DP)?

In General

- DP links amount of randomization to risk thresholds.
- DP calls for a rigorous tracking of all data uses and releases.
- DP does not make explicit assumptions about the knowledge and behavior of the intruder.

For Our Discussion

- Additivity of risk across releases based on worst case scenario .
- Room for improvement, especially if worst case is ∞ risk
- Room for improvement, especially if some knowledge is considered public.

Outline

Differential Privacy under $\mathcal{M} = \xi(\theta \mid \mathbf{x})$

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

- ϵ bounds the **change** in the probability measure ξ
  - from the inclusion of a **single record** $\delta(\mathbf{x}, \mathbf{x}^{'}) = 1$,
  - over **all possible outcomes**, $B \in \beta_\Theta$ – sets in the space of measurable sets of Θ.
  - over **all possible data sets** $\mathbf{x}, \mathbf{x}^{'} \in \mathcal{X}^n$ of size n.

DP under the Posterior Mechanism

Dimitrakakis et al. (2017) establish a link between bounding the log-likelihood $f_\theta(\mathbf{x}) = \log \pi_\theta(\mathbf{x})$ and a DP bound ϵ.

- If $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}^n : \delta(\mathbf{x},\mathbf{x}')=1} \sup_{\theta \in \Theta} |f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}')| \leq \Delta$.
- Then a single draw of replicate data $\mathbf{x}^* \sim \pi(\mathbf{x}^*|\mathbf{x})$ has a DP guarantee of $\epsilon \leq 2\Delta$.

Major Limitations of the Posterior Mechanism

The results in Dimitrakakis et al. (2017) are not immediately practical.

- For many data distributions of interest (e.g. normal, exponential, Poisson, geometric), $\Delta = \infty$.
  - Directly truncating the support Θ and the domain $\mathcal{X}^n$ may work for simple distributions but is somewhat ad-hoc and scales poorly with the dimension of both
- Using a prior that induces more smoothing may reduce Δ – but requires re-estimation. This is an indirect adjustment.

Pseudo Posterior Mechanism

- Savitsky et al. (2022) utilize record-indexed weights, $\alpha \in (0, 1]^n$
- To downweight likelihood contributions with high disclosure risk

$$\xi^{\boldsymbol{\alpha}}\left(\theta \mid \mathbf{x}, \gamma\right) \propto \left[\prod_{i=1}^{n} \pi\left(x_i \mid \theta\right)^{\alpha_i}\right] \pi\left(\theta \mid \gamma\right)$$

- $\alpha_i \propto 1/\sup_{\theta \in \Theta} \left| f_\theta\left(x_i\right) \right|$
- Allows surgical downweighting of high risk records
- α_i induces an anti-informative prior
- $\sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{x}') = 1} \sup_{\theta \in \Theta} |\alpha(\mathbf{x}) f_\theta(\mathbf{x}) - \alpha(\mathbf{x}') f_\theta(\mathbf{x}')| \leq \Delta_\alpha$
- $\Delta_\alpha \leq \Delta$
- Ensures $\Delta_\alpha < \infty$
- Expected to better preserve real data distribution for any target privacy budget, ϵ

Outline

(A) Measure Expert Beliefs as modeled Probability

- Expert believes interested public already knows that the true datum value does not lie in the space outside of the ball/interval
- $p_{\mathcal{M}}(\mathbf{x}^* \mid \mathbf{x})$, constructed from the same model estimated on the first step to construct the risk-based weights, α_i.

$$\lambda_i = \Pr_{\mathcal{M},a,b}\left(x_i^* \notin [a \times x_i, b \times x_i]\right) = \Pr_{\mathcal{M},R_i}\left(x_i^* \notin R_i\right)$$

- Each λ_i, then, represents the probability that a realized datum value for unit i does *not* need to be protected

(A) Risk-based Weights Adjusted by Expert Beliefs

$$\alpha_i^* = \lambda_i + (1 - \lambda_i) \times \alpha_i$$

- When a unit is highly risky such that $\alpha_i = 0$ (is set to 0) then $\alpha_i^* = \lambda_i$.
- By contrast, when datum i presents no disclosure risk such that $\alpha_i = 1$, then $\alpha_i^* = 1$, indicating there is no privacy protection required under this scenario.
- When $\lambda_i = 0$ it is supposed that the interested public has *no* knowledge about a subspace of the private variable and we return to the usual case where $\alpha_i^* = \alpha_i$

(A) Reduced Sensitivity by Incorporating Probabilities

$$p_{\boldsymbol{\theta}}^{\boldsymbol{\alpha}*}(\mathbf{x}) = p_{\boldsymbol{\theta}}^{\boldsymbol{\lambda}^c\boldsymbol{\alpha}}(\mathbf{x}) p_{\boldsymbol{\theta}}^{\boldsymbol{\lambda}}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})^{(1-\lambda_i)\alpha_i} \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})^{\lambda_i}.$$

$$\begin{aligned}\Delta_{\boldsymbol{\alpha},\boldsymbol{\lambda},\mathbf{x}} &= \max_{\theta \in \{\xi^{\boldsymbol{\alpha}^*(\boldsymbol{x})}(\theta|\boldsymbol{x})\}_m} \max_{i \in 1,\ldots,n} |\alpha_i^* \times f_{\theta_m}(x_i) - \lambda_i \times f_{\theta_m}(x_i)| \\ &= \max_{\theta \in \{\xi^{\boldsymbol{\alpha}^*(\boldsymbol{x})}(\theta|\boldsymbol{x})\}_m} \max_{i \in 1,\ldots,n} |\big((1-\lambda_i)\alpha_i\big) \times f_{\theta_m}(x_i)| \leq \Delta_{\boldsymbol{\alpha},\mathbf{x}},\end{aligned}$$

- Further reduce the sensitivity by factor of $(1 - \lambda_i)$.
- The $(1 - \lambda_i)$ term represents the portion of the data support that is *not* known by the interested public and, therefore, sensitive.
- Under our set-up, we account for expert beliefs *indirectly* through knowledge weights $\boldsymbol{\lambda}$ rather than directly truncating the data support, $\mathcal{X}^n$.

(B) Minimal Use of Expert Knowledge

- The specification of the range $R_i = [a \times x_i, b \times x_i]$ gives rise to
  - the truncation adjustment $P_\theta(R_i) = \int_{x \in R_i} p_\theta(x) dx$
  - the range-truncated likelihood $p_{\theta_i}(x_i)/P_\theta(R_i)$
- Assuming the range R_i is public, we can use R_i for 'free' in inference, analogous to using censoring times in survival analysis.

$$\begin{aligned} p^{I^c,\alpha_i}(x_i|\theta,a,b) = & \quad p(x_i|\theta)^{\alpha_i} / \left(P(b \times x_i|\theta) - P(a \times x_i|\theta)\right) \\ = & \quad p(x_i|\theta)^{\alpha_i} / p^I(x_i|\theta,a,b) \end{aligned}$$

(B) Reduced Sensitivity by Incorporating End Points

$$p_{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}(\mathbf{x}) = p_{\boldsymbol{\theta}}^{\boldsymbol{I}^c\boldsymbol{\alpha}}(\mathbf{x}) p_{\boldsymbol{\theta}}^{\boldsymbol{I}}(\mathbf{x}) = \prod_{i=1}^{n} p^{I^c,\alpha_i}(x_i \mid \theta, a, b) \prod_{i=1}^{n} p^{I}(x_i \mid \theta, a, b).$$

$$\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}} = \max_{\theta \in \{\xi(\theta|\boldsymbol{x})\}_m} \max_{i \in 1,\ldots,n} |\alpha_i f_{\theta_m}(x_i) - \log\big(P_{\theta_m}(R_i)\big)| \leq \Delta_{\boldsymbol{\alpha},\mathbf{x}}$$

- Further reduce the risk-based weighted sensitivity by

$$\log\big(P_{\theta_m}(R_i)\big) = \log\big(P(b \times x_i|\theta_m) - P(a \times x_i|\theta_m)\big)$$

- Rather than directly truncating the data support, $\mathcal{X}^n$ and greatly increasing the complexity of the synthesizer $p_{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}(\mathbf{x})$, we partition the sensitivity calculation into contributions coming from within and outside the restricted range.

Range Restricted Privacy under the Pseudo Posterior

$$(A) \quad \sup_{\mathbf{x}\in\mathcal{X}^n,\mathbf{x}'\in\mathcal{X}^{n-1}:\delta(\mathbf{x},\mathbf{x}')=1} \sup_{B\in\beta_\Theta} \frac{\xi^{\boldsymbol{\lambda}^c\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x},\mathbf{R})}{\xi^{\boldsymbol{\lambda}^c\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x},\mathbf{R})} \leq e^{\epsilon},$$

$$(B) \quad \sup_{\mathbf{x}\in\mathcal{X}^n,\mathbf{x}'\in\mathcal{X}^{n-1}:\delta(\mathbf{x},\mathbf{x}')=1} \sup_{B\in\beta_\Theta} \frac{\xi^{\boldsymbol{I}^c\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x},\mathbf{R})}{\xi^{\boldsymbol{I}^c\boldsymbol{\alpha}(\mathbf{x}')}(B \mid \mathbf{x}',\mathbf{R})} \leq e^{\epsilon},$$

Sensitivities under Intruder Knowledge

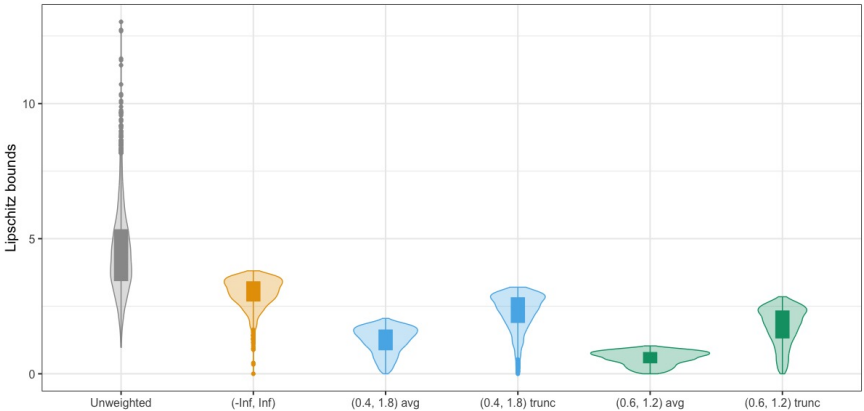


Figure: Violin plots of by-record Lipschitz bounds of Unweighted, (-Inf, Inf) (i.e., no bounds as Weighted), $(0.4, 1.8)$ averaged, $(0.4, 1.8)$ truncated, $(0.6, 1.2)$ averaged, and $(0.6, 1.2)$ truncated, over a single sample.

Preservation of Quantiles under MC study

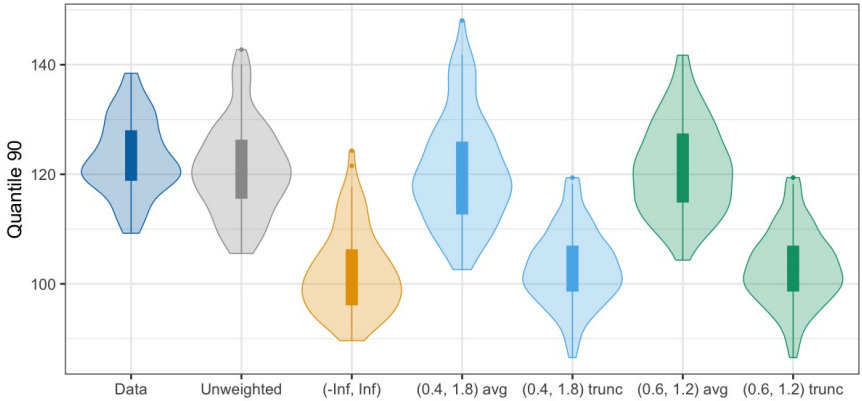


Figure: Violin plots of Q90s of Unweighted, (-Inf, Inf) (i.e., no bounds as Weighted), $(0.4, 1.8)$ averaged, $(0.4, 1.8)$ truncated, $(0.6, 1.2)$ averaged, and $(0.6, 1.2)$ truncated, over 100 repeated samples.

Real-vs-Synthetic Distribution Distance under MC study

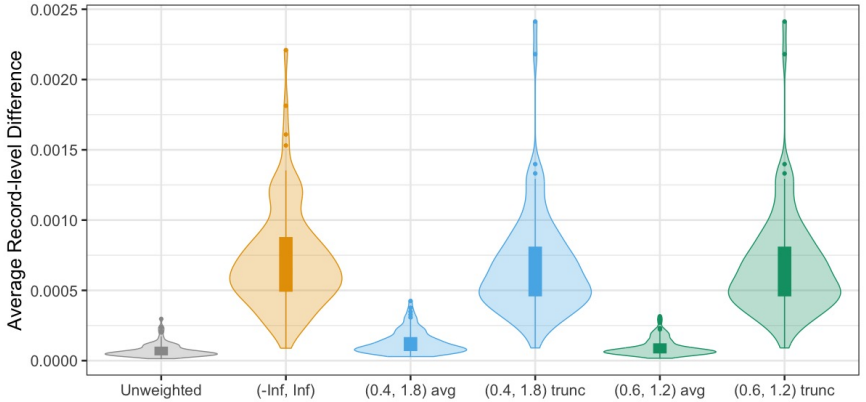
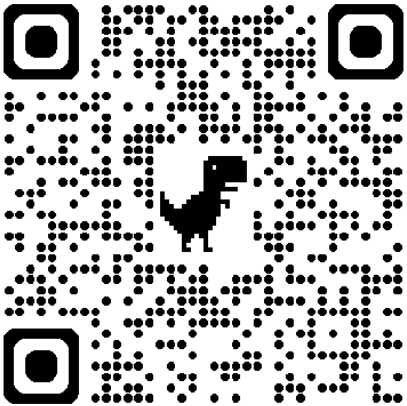


Figure: Violin plots of average metric of ECDF of Unweighted, (-Inf, Inf) (i.e., no bounds as Weighted), $(0.4, 1.8)$ averaged, $(0.4, 1.8)$ truncated, $(0.6, 1.2)$ averaged, and $(0.6, 1.2)$ truncated, over 100 repeated samples.

Preprint

https://arxiv.org/abs/2602.04124



References I

Dimitrakakis, C., Nelson, B., Zhang, Z., Mitrokotsa, A. and Rubinstein, B. I. P. (2017), 'Differential privacy for bayesian inference through posterior sampling', *Journal of Machine Learning Research* **18**(1), 343–381.

Hu, J. (2019), 'Bayesian estimation of attribute and identification disclosure risks in synthetic data', *Transactions on Data Privacy* **12**, 61–89.

Savitsky, T. D., Williams, M. R. and Hu, J. (2022), 'Bayesian pseudo posterior mechanism under asymptotic differential privacy', *Journal of Machine Learning Research* **23**, 1–37.

Extra Slides

Local vs. Global Privacy Guarantee

When we implement

- We have to estimate $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ based on a single data set, $\mathbf{x}$, to estimate the DP guarantee ϵ. (local DP result)
- We have to approximate the $\sup_{\theta\in\Theta} f_\theta(x_i)$ as $\max_{\theta_j, j\in 1,\ldots,J} f_{\theta_j}(x_i)$.
- Overestimate $\boldsymbol{\alpha} \rightarrow$ Underestimate ϵ

Asymptotic Global Guarantee

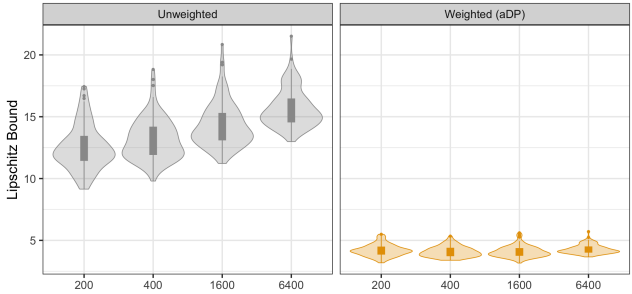
To justify a global DP result (bounding all data sets) compared to a local DP result (bound observed data set):

Asymptotic "Discovery" of $\Delta_{\boldsymbol{\alpha}}$ at large sample sizes (n)

- Space of plausible values Θ collapses to a point θ^*, so don't need to look at $\sup_{\theta \in \Theta}$.
- Variation across local $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ collapses onto $\Delta_{\boldsymbol{\alpha}}$.
- Achieves $(\epsilon, \delta)-$ pDP, where $\delta > 0$ is probability $\exists \mathbf{x} \in \mathcal{X}^n$ exceeding the ϵ bound.
  - $\delta \to 0$ at $\mathcal{O}(n^{-1/2})$.
- Requires increasing sparsity in downweighted record contributions, which aligns with focus on isolated records as risky.

Contraction of Lipschitz, $(\Delta_{\boldsymbol{\alpha},\mathbf{x}_r})$

- Generate $x_{ri} \overset{\text{ind}}{\sim}$ lognormal for $r = 100$.



Asymptotic Contraction of Local Privacy Guarantee

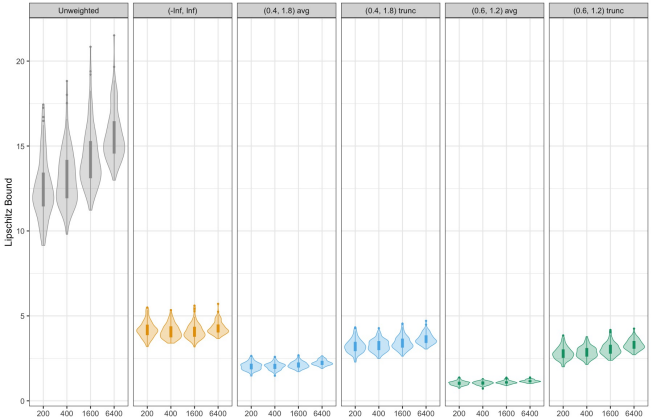


Figure: Violin plots of Lipschitz bounds of Unweighted, (-Inf, Inf) (i.e., no bounds as Weighted), $(0.4, 1.8)$ averaged, $(0.4, 1.8)$ truncated, $(0.6, 1.2)$ averaged, and $(0.6, 1.2)$ truncated, over 100 repeated samples and four different sample sizes $n = \{200, 400, 1600, 6400\}$.