PRIVACY AMPLIFICATION FOR SYNTHETIC DATA USING RANGE RESTRICTION (PART 2)

Jingchen Hu, Matthew R. Williams, Terrance D. Savitsky

Binghamton University, RTI International, U.S. Bureau of Labor Statistics

We introduce a new class of range restricted formal data privacy standards that condition on owner beliefs about sensitive data ranges. By incorporating this additional information, we can provide a stronger privacy guarantee (e.g., an amplification). The range restricted formal privacy standards protect only a subset of data values and exclude rangesbelieved to be already publicly known. The privacy standards are designed for the risk-weighted pseudo posterior (model) mechanism (PPM) used to generate synthetic data under an asymptotic differential (aDP) privacy guarantee [1]. We compare two alternative adjustments. The first expresses data owner knowledge of the sensitive range as a probability, λ, that a datum value drawn from the underlying generating distribution lies *outside* the ball or subspace of values that are sensitive. The second adjustment encodes knowledge as the difference in probability masses $P(R) \leq 1$ between the edges of the sensitive range, R. We use the resulting *conditional* (pseudo) likelihood for a sensitive record, which boosts its worst case tail values away from 0.

Suppose the owner of the closely-held data defines a width with respect to sensitive variable, $\mathbf{x}$, as a percentile of the distribution for each $x_i,\ i \in (1, \ldots, n)$. We construct $r = [a, b]$ and state that the interested public knows that the variable value for record i lies inside of the interval $[a \times x_i, b \times x_i]$; i.e., we propose to restrict privacy protection to the portion of the data support *inside* the interval of interest where the intruder does not know where lies the true value, x_i, for record i. More formally, we could measure the known information on sensitive ranges by constructing the following event probability, $\lambda_i = \Pr\left(x_i^* \notin [a \times x_i, b \times x_i]\right)$ from the posterior predictive distribution. Part 1 discusses the formation of a privacy standard and resulting privacy properties of incorporating this information for x_i.

In Part 2, we present a more conservative approach using the "minimum" information about the sensitive range. We declare the datum for a record as either lying inside or outside of that range and condition our use of the model distribution only on endpoints of the sensitive interval (rather than the distribution of data values within the sensitive range). For an arbitrary interval $[a \times x_i, b \times x_i] \in \mathcal{X}$, we propose an interval censored formulation for the assumed known likelihood component as

$$p^I(x_i|\theta, a, b) = \begin{cases} P(b \times x_i|\theta) - P(a \times x_i|\theta), & x_i \in [a \times x_i, b \times x_i] \\ p(x_i|\theta), & x \notin [a \times x_i, b \times x_i] \end{cases}$$

with $P(\cdot|\theta)$ being the cdf under our model for the closely-held data. This likelihood represents information generally known because when x_i is in the sensitive range the edges of that range are known. By contrast, when x_i is not in the sensitive range there is no need to privatize. The specification of this range gives rise to the truncation adjustment $P_\theta(R_i) = \int_{x \in R_i} p_\theta(x)dx$ and the range-truncated likelihood: $p_{\theta_i}(x_i)/P_\theta(R_i)$.

Definition 1 *(Range-Truncated Privacy under the Posterior Mechanism)*

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

For a database sequence, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ under $x_1, \ldots, x_n \sim P_{\theta_0}$, for some $\theta_0 \in \Theta$, we formulate the truncated (pseudo-) likelihood,

$$p_\theta^{\boldsymbol{I}^c}(\mathbf{x}) = \prod_{i=1}^n p_\theta(x_i)/P_\theta(R_i), \tag{1}$$

for each $\theta \in \Theta$ and $\mathbf{x} \in \mathcal{X}^n$.

$$\xi^{\boldsymbol{I}^c(\mathbf{x})}(B \mid \mathbf{x}) = \frac{\int_{\theta \in B} p_\theta^{\boldsymbol{I}^c}(\mathbf{x}) d\xi(\theta)(\mathbf{x})}{\phi^{\boldsymbol{I}^c(\mathbf{x})}}, \tag{2}$$

where $\phi^{\boldsymbol{I}^c(\mathbf{x})}(\mathbf{x}) \triangleq \int_{\theta \in \Theta} p_\theta^{\boldsymbol{I}^c}(\mathbf{x}) d\xi(\theta)$ normalizes the pseudo posterior distribution.

We prove that the risk-weighted pseudo posterior $\xi^{\boldsymbol{I}^c \boldsymbol{\alpha}(\mathbf{x})}(\cdot \mid \mathbf{x})$ has local privacy guarantee $2\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}}$ with sensitivity

$$\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}} = \max_{\theta \in \{\xi^{\boldsymbol{\alpha}^*(\boldsymbol{x})}(\theta|\boldsymbol{x})\}_m} \max_{i \in 1,\ldots,n} |\alpha_i \times f_{\theta_m}(x_i) - \log(P(b \times x_i|\theta_m) - P(a \times x_i|\theta_m))|. \tag{3}$$

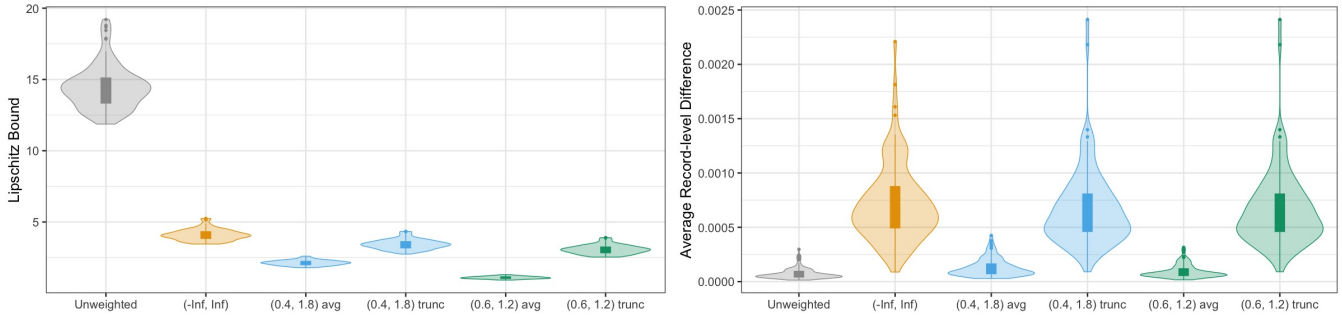
Theorem 1 $\forall \mathbf{x} \in \mathcal{X}^n, \mathbf{x} \in \mathcal{X}^{n-1} : \delta(\mathbf{x}, \mathbf{x}) = 1, B \in \beta_\Theta$ *(where* β_Θ *is the* $\sigma-$*algebra of measurable sets on* Θ*) under* $\boldsymbol{\alpha}(\cdot)$ *with* $\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}} > 0$*,*

$$\sup_{B \in \beta_\Theta} \frac{\xi^{\boldsymbol{I}^c \boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x})}{\xi^{\boldsymbol{I}^c \boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x})} \leq \exp(2\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}}), \tag{4}$$

i.e. the pseudo posterior $\xi^{\boldsymbol{I}^c \boldsymbol{\alpha}(\mathbf{x})}(\cdot \mid \mathbf{x})$ *has local privacy guarantee* $2\Delta_{\boldsymbol{\alpha},\boldsymbol{I},\mathbf{x}}$*.*

In each repeated simulation runs, we simulate $n = 2000$ records from $x \sim \text{Normal}(2, 1)$ and $y \sim \text{Lognormal}(x + 1, 1)$ in order to mimic highly skewed data in real applications. We use $(a, b) = \{(0.4, 1.8), (0.6, 1.2)\}$.

Figure 1a shows the privacy budget (twice of the maximum Lipschitz bound) decreases under sensitive range for both range-averaged and range-truncated standards. With the same (a, b) choice, the range-averaged standard produces further privacy budget decrease compared to the range-truncated standard. Utility results on the ECDF statistics in Figure 1b shows that the range-averaged standard preserves higher utility compared to the range-truncated standard with the same (a, b) choice.



(a) Lipschitz bounds, 100 repeated samples **(b)** ECDF

Figure 1. Violin plots of (a) Lipschitz bounds and (b) ECDF utility results.

Our simulation results also confirm that incorporating a sensitive range maintains the asymptotic feature of aDP that as the sample size increases, the local $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ contracts onto the global $\Delta_{\boldsymbol{\alpha}}$ (figure omitted for brevity).

References

[1] T. D. Savitsky, M. R. Williams, and J. Hu. "Bayesian pseudo posterior mechanism under asymptotic differential privacy". In: *Journal of Machine Learning Research* 23 (2022), pp. 1–37.