CERIAS

The Center for Education and Research in Information Assurance and Security

# Semi Differential Privacy

*Young Hyun Cho and Jordan Awan*
*Department of Statistics, Purdue University*

## Motivation

### Differential Privacy and the 2020 Census

**How will differential privacy affect 2020 census data?**

-Differential privacy has strong guarantees, but is not always implemented exactly

-For example, US Census publishes a combination of DP and exact statistics

-Currently NO satisfactory method of quantifying the privacy in such settings

## Preliminary

A randomized mechanism $M(X)$ is $(\varepsilon, \delta)$-Differential Privacy if for any $\varepsilon, \delta > 0$,

$$\mathbb{P}(M(X) \in S) \leq e^{\varepsilon}\mathbb{P}(M(X') \in S) + \delta,$$

$\rho - zCDP$ if for any $\rho \geq 0$ and for all $\alpha \in (1, \infty)$,

$$D_{\alpha}(M(X)||M(X')) \leq \rho\alpha,$$

where $D_{\alpha}$ is $\alpha$-Renyi divergence,
for any databases X,X' such that d(X,X') ≤ 1

Smaller Privacy Parameters $(\varepsilon, \delta)$ $or$ $\rho \Longrightarrow$ Stronger Privacy

BUT THE SAME SPIRIT:
Any Pair of Input Data X,X' with ONE DIFFERENT ENTRY is Difficult to distinguish by observing the output of $M$.

**-What happens if true statistic T(X) is published along with DP M(X)?**
**-Especially, what if there is no neighboring data X,X' s.t. T(X) = T(X')? That is,**

$$\{(\boldsymbol{X}, \boldsymbol{X}'): \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{X}') \text{ } \boldsymbol{and} \text{ } \boldsymbol{d}(\boldsymbol{X}, \boldsymbol{X}') \leq \mathbf{1}\} = \boldsymbol{\phi}$$

We propose a framework semi-DP, which properly accounts for the combination of private and non-private releases.
Under semi-DP we derive optimal mechanisms under a variety of scenarios.

## Definition of Semi-DP

**Algorithm 1** Semi-adjacent Parameter $a$

**Input:** True Statistics $T(X)$
Define $\mathcal{X}^n|_T = \{X' \in \mathcal{X}^n : T(X') = T(X)\}$
**for** $i \in [n]$ **do**
  Define $\mathcal{P}_i' = \{x_i' \in \mathcal{X} : \exists Y \in \mathcal{X}^{n-1} \text{ s.t. } T\left(\begin{bmatrix} x_i' \\ Y \end{bmatrix}\right) = T\left(\begin{bmatrix} x_i \\ X_{-i} \end{bmatrix}\right)\}$
  **if** $\mathcal{P}_i' = \emptyset$ **then**
    Put $a_i = \infty$
  **else**
    Define $\mathcal{X}_i' = \left\{\left(\begin{bmatrix} x_i' \\ Y \end{bmatrix}\right) : x_i' \in \mathcal{P}_i', Y = \text{argmin}_{Y \in \mathcal{X}^{n-1}} d(X_{-i}, Y) \text{ s.t. } T\left(\begin{bmatrix} x_i' \\ Y \end{bmatrix}\right) = T(X)\right\}$
    $a_i = \max_{X_i' \in \mathcal{X}_i'} d(X, X_i')$
  **end if**
**end for**
**Output:** $a = \max_{i \in [n]} a_i$

Given semi-adjacent parameter $\boldsymbol{a}$, our semi-DP makes difficult to distinguish any pair of data in

$$\{(\boldsymbol{X}, \boldsymbol{X}'): \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{X}') \text{ } \boldsymbol{and} \text{ } \boldsymbol{d}(\boldsymbol{X}, \boldsymbol{X}') \leq \boldsymbol{a}\}$$

## Analysis on US 2020 Census

T(X): True Population for each state, the District of Columbia and Puerto Rico
$\Longrightarrow$ semi-adjacent parameter $\boldsymbol{a} = \mathbf{2}$

Census Advertised: $\rho$-zCDP with $\rho = 0.213$
Semi-DP framework: $\rho$-zCDP with $\rho = 0.852$

## Mechanism Design

Goal: Add the least noise and achieve the target privacy parameter

**Algorithm 2** `Generalized Optimal` $K$`-norm mechanism:`$\epsilon$`-semi DP`

**Input:** Statistics $T(X)$ to be released, Mechanism $\theta$ to be privatized
Calculate the semi-adjacent parameter $a$
Define the sensitivity space $S_{\theta|T} = \{\theta(X) - \theta(X') : d(X, X') \leq a, T(X) = T(X')\}$
Define $\text{Proj}_{S_{\theta|T}}$ be the orthogonal projection operator on to span$(S_{\theta|T})$
Calculate $\text{S}_{\text{Proj}} = \{\text{Proj}_{S_{\theta|T}}\theta(X) - \text{Proj}_{S_{\theta|T}}\theta(X') : d(X, X') \leq a, T(X) = T(X')\}$
Let Hull$(S_{\text{Proj}})$ be a convex hull of $S_{\text{Proj}}$
and define a norm $\|\cdot\|_K$ by $\|u\|_K = \inf\{c \in \mathbb{R}^{\geq 0} : u \in cH\}$
Calculate $K$-norm sensitivity $\Delta_K = \sup_{u \in S_{\text{Proj}}} \|u\|_K$
Sample $V \sim f_V(v) = c \exp\left(-\frac{\epsilon}{\Delta_K}\|v\|_K\right)$
**Output:** $M(X) = \theta(X) + V$

Generalized Optimal K-norm mechanism achieves $\varepsilon$-semi DP with the least additive noise, so optimal!

**Algorithm 3** `Gaussian Mechanism:`$\mu$`-semi GDP`

**Input:** Statistics $T(X)$ to be released, Mechanism $\theta$ to be privatized
Calculate the semi-adjacent parameter $a$
Define the sensitivity space $S_{\theta|T} = \{\theta(X) - \theta(X') : d(X, X') \leq a, T(X) = T(X')\}$
Calculate $P$ be the orthogonal projection matrix onto span$(S_{\theta|T})$
Calculate $l_2$-norm sensitivity $\Delta_2 = \sup_{u \in S_{\theta|T}} \|u\|_K$
Sample $N \sim N(0, (\frac{\Delta_2}{\mu})^2 P)$
**Output:** $M(X) = \theta(X) + N$

Gaussian Mechanism achieves $\mu$-semi GDP with less noise than naïve approaches
Additive Gaussian noise can also achieve $(\varepsilon, \delta)$-DP and $\rho$-zCDP

PURDUE UNIVERSITY

CERIAS