Post-processing of Differentially Private Data: A Fairness Perspective

Keyu Zhu[1] , **Ferdinando Fioretto**[2] and **Pascal Van Hentenryck**[1]
[1]Georgia Institute of Technology
[2]Syracuse University
kzhu67@gatech.edu, ffiorett@syr.edu, pvh@isye.gatech.edu

Abstract

Post-processing immunity is a fundamental property of differential privacy: it enables arbitrary data-independent transformations to differentially private outputs without affecting their privacy guarantees. Post-processing is routinely applied in data-release applications, including census data, which are then used to make allocations with substantial societal impacts. This paper shows that post-processing causes disparate impacts on individuals or groups and analyzes two critical settings: the release of differentially private datasets and the use of such private datasets for downstream decisions, such as the allocation of funds informed by US Census data. In the first setting, the paper proposes tight bounds on the unfairness for traditional post-processing mechanisms, giving a unique tool to decision makers to quantify the disparate impacts introduced by their release. In the second setting, this paper proposes a novel post-processing mechanism that is (approximately) optimal under different fairness metrics, either reducing fairness issues substantially or reducing the cost of privacy. The theoretical analysis is complemented with numerical simulations on Census data.

1 Introduction

Differential privacy (DP) [Dwork *et al.*, 2006] has become a fundamental technology for private data release. Private companies and federal agencies are rapidly developing their own implementations of DP. It is particularly significant to note that the U.S. Census Bureau adopted DP for its 2020 release [Abowd, 2018]. It is also of primary importance to observe that the released data by corporation or federal agencies are often used to make policy decisions with significant societal and economic impacts for the involved individuals. For example, U.S. census data users rely on the decennial census data to apportion the 435 congressional seats, allocate the $1.5 trillion budget, and distribute critical resources to U.S. states and jurisdictions.

Although DP provides strong privacy guarantees on the released data and is widely celebrated among privacy researchers, its wide adoption among more federal agencies and public policy makers presents a key challenge: without careful considerations, DP methods may disproportionately impact minorities in decision processes based on the private data. Specifically, to protect individuals in a dataset, typical DP data-release methods operate by adding calibrated noise onto the data and then *post-process* the resulting noisy data to restore some important data invariants. Since such a process perturbs the original data, it necessarily introduces some errors which propagate onto downstream decision tasks. In fact, this paper will show that these errors may affect various individuals differently. Although understanding the outcome of these effects is extremely important, these disproportionate impacts are poorly understood and have not received the attention they deserve given their broad impact on various population segments.

This paper addresses this gap in understanding the effect of DP, and analyzes the disproportionate effects of a family of post-processing methods commonly adopted in data release tasks. The analysis focuses on two critical settings: the release of differentially private datasets and the use of such private datasets in critical allocation tasks, as those using U.S. Census data to allocate funds and benefits. The paper makes two fundamental contributions:

1. In the release setting, the paper derives tight bounds on the unfairness introduced by commonly adopted post-processing mechanisms, providing a valuable tool for policy makers and information officers to understand the disproportionate impact of their DP releases. These results are complemented by numerical simulations on the census data.
2. In the downstream decision setting, the paper proposes a novel post-processing mechanism that integrates the data invariant into the downstream decision processes. The resulting mechanism achieves near-optimal results and reduces unfairness and the cost of privacy up to an order of magnitude on practical case studies.

To the best of the authors' knowledge, this is the first study that analyzes the fairness impacts of DP post-processing steps. The rest of this paper presents the related work, the preliminaries, the settings considered, and the motivation. The core of the paper is in Sections 6 and 7 that present the two main contributions. The last section concludes the paper. All the proofs are in the Appendices that also contain a nomenclature summary.

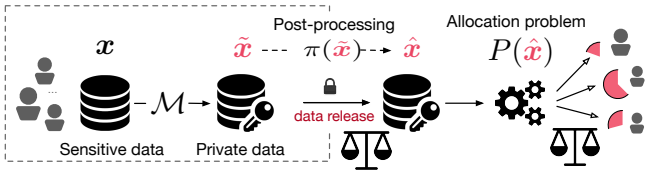

Figure 1: Schematic problem representation.

2 Related Work

Privacy and fairness have been studied mostly in isolation with a few exceptions. [Cummings *et al.*, 2019] considered the tradeoffs arising between differential privacy and equal opportunity. [Ekstrand *et al.*, 2018] raised questions about the tradeoffs involved between privacy and fairness, and [Jagielski *et al.*, 2018] showed two algorithms that satisfy (ϵ, δ)-differential privacy and equalized odds. In the context of data release and resource allocation, [Pujol *et al.*, 2020] were seemingly first to show, empirically, that there might be privacy-fairness tradeoffs involved in resource allocation settings. In particular, for census data, they show that the noise added to achieve differential privacy could disproportionately affect some groups over others. [Tran *et al.*, 2021] formalized the ideas developed in [Pujol *et al.*, 2020] and characterized the conditions for which fairness violations can be bounded for a class of allocation problems. Finally, [Abowd and Schmutte, 2019] considered statistical accuracy and privacy protection as competing public goods, and designed an economic framework to balance the tradeoff. *This paper departs from these results significantly: it provides tight lower and upper bounds on the unfairness introduced by post-processing steps that are critical for practical applications, and proposes new mechanisms that merge post-processing and the downstream resource allocation for mitigating these fairness issues.*

3 Preliminaries: Differential Privacy

Differential Privacy [Dwork *et al.*, 2006] (DP) characterizes the amount of individual data disclosed in a computation.

Definition 1. *A randomized algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies* (ϵ, δ)-differential privacy *if for any output $O \subseteq \mathcal{R}$ and datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ differing by at most one entry (written as $\boldsymbol{x} \sim \boldsymbol{x}'$)*

$$\Pr[\mathcal{M}(\boldsymbol{x}) \in O] \leqslant \exp(\epsilon) \Pr[\mathcal{M}(\boldsymbol{x}') \in O] + \delta. \quad (1)$$

Parameter $\epsilon > 0$ is the *privacy loss*: values close to 0 denote strong privacy and $\delta \geqslant 0$ represents a probability of failure. Intuitively, DP states that every event has a similar probability regardless of the participation of any individual data to the dataset. DP satisfies several properties including *immunity to post-processing*, which states that the privacy loss of DP outputs is not affected by arbitrary data-independent post-processing [Dwork and Roth, 2013].

A function f from a dataset $\boldsymbol{x} \in \mathcal{X}$ to a result set $R \subseteq \mathbb{R}^n$ can be made differentially private by injecting random noise onto its output. The amount of noise relies on the notion of *global sensitivity* $\Delta_f = \max_{\boldsymbol{x} \sim \boldsymbol{x}'} \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_p$ with $p \in \{1, 2\}$. The *Laplace mechanism* [Dwork *et al.*, 2006] that outputs $f(\boldsymbol{x}) + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^n$ is drawn from the i.i.d. Laplace distribution with 0 mean and scale Δ_f/ϵ over n dimensions, achieves ϵ-DP. The *Gaussian mechanism* [Dwork and Roth, 2013] that outputs $f(D) + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^n$ is drawn from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ with parameter $\sigma \geqslant c\Delta_f/\epsilon$, achieves (ϵ, δ)-differential privacy, for $c^2 > 2\ln(1.25/\delta)$.

4 Settings and Goals

The paper considers datasets $\boldsymbol{x} \in \mathbb{R}$ of n entities, whose elements x_i describe some measurable quantities of entity $i \in [n]$, such as the number of individuals living in a geographical region i. A data-release mechanism $\mathcal{M}$ is applied to the dataset $\boldsymbol{x}$ (called true data in this paper) to produce a privacy-preserving counterpart $\tilde{\boldsymbol{x}} \sim \mathcal{M}(\boldsymbol{x})$ (referred to as noisy data). Given the released data, the paper considers allocation problems $P : \mathbb{R}^n \to \mathbb{R}^n$ that distribute a finite set of resources to the problem entities. For example, P may be used to allocate funds to school districts.

The focus of the paper is to study the error disparities of a DP data-release mechanism $\mathcal{M}$ in two contexts: **(1)** data release and **(2)** downstream decisions. The first context refers to the case in which the noisy data must be post-processed before being released to satisfy desired invariants. The second context refers to the case in which the noisy data is released for use in an allocation problem. Again, the release data must be post-processed to satisfy the problem-specific feasibility constraints. *The paper studies the disparate impacts of the error introduced by post-processing among entities in both scenarios.*

Quantitatively, this error is represented by the bias associated with a post-processing mechanism π, i.e.,

$$\mathcal{B}(\pi, P, \mathcal{M}, \boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{M}(\boldsymbol{x})}[\pi(\tilde{\boldsymbol{x}})] - P(\boldsymbol{x}).$$

The paper will often omit the last two arguments of the bias term when there is no ambiguity. The disparate impact of the error is then characterized by the following definition.

Definition 2 (α-fairness)**.** *A post-processing mechanism π is said α-fair with respect to problem P if the maximum difference among the biases is bounded by α, i.e.,*

$$\|\mathcal{B}(\pi, P)\|_{\rightleftharpoons} = \max_{i \in [n]} \mathcal{B}(\pi, P)_i - \min_{i \in [n]} \mathcal{B}(\pi, P)_i \leqslant \alpha$$

with α referred to as a *fairness bound* that captures the fairness violation.

5 Motivating Applications

This section reviews two settings highlighting the disparate impacts of DP post-processing in census releases.

Data Release. Consider a simplified version of the census data release problem. The task is to release counts, such as demographic information of individuals, which are required to be non-negative and summed up to a public quantity. The latter is usually used to preserve known statistics at a state or national level. To preserve these invariants, commonly adopted post-processing mechanisms (e.g., the one

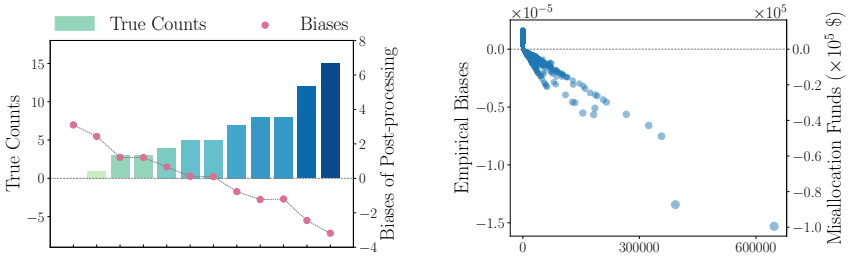

Figure 2: **Data Release**: (Left) bar chart of the true counts and line chart of the empirical biases (red dots) associated with the given post-processing mechanism. **Downstream decisions**: (Right) scatter plot of the empirical biases (left y-axis) and misallocation funds (right y-axis) resulting from the given post-processing mechanism against different school district sizes (x-axis). For both instances, Laplace mechanism is taken for privacy protection and each experiment is repeated for 200,000 times.

used by the Top-Down algorithm in the release of several 2020 U.S. census data products) constrain the noisy DP outcomes with an ℓ_2 *projection* step. Such post-processing step will be described in detail and studied in the next section. Figure 2 (left) shows the (sorted) values of some synthetically generated true counts (bars) and the (empirical) biases (red dots), obtained by comparing the post-processed DP counts with the true counterparts. Notice how the resulting biases vary among entities. *This is significant as sub-communities may be substantially under- or over-counted affecting some important data analysis tasks.*

Downstream Decisions. These disparities may also have negative socio-economic impacts. For instance, when agencies allocate funds and benefits according to differentially private data, an ill-chosen post-processing mechanism can result in huge disparities and, as a consequence, lead to significant inefficiencies of allocation. Consider the *Title I of the Elementary and Secondary Education Act of 1965* [Sonnenberg, 2016]: It uses the US Census data to distribute about $6.5 billion in basic grants to qualified school districts in proportion to the count x_i of children aged 5 to 17 who live in necessitous families in district i. The allocation is formalized by

$$P_i^F(\boldsymbol{x}) := \frac{a_i \cdot x_i}{\sum_{j=1}^n a_j \cdot x_j}, \qquad \forall\, i \in [n],$$

where $\boldsymbol{x} = [x_1 \ \dots \ x_n]^\top$ is the vector of the districts' true counts and a_i is a positive weight factor reflecting students expenditures in district i. When a projection mechanism (described in more details in Section 7) is used to guarantee non-negativity of the private data $\tilde{\boldsymbol{x}}$, the resulting errors on the proposal of funds allocation can be consequential. Figure 2 (right) visualizes the misallocation (blue dots) for over 16,000 school districts (due to this post-processing mechanism) in terms of proportions (left y-axis) and funds (right y-axis). In this numerical simulation, which uses data based on the 2010 US census release, *large school districts* may be strongly penalized. For example the largest school district in Los Angeles can receive up to 99,000 dollars fewer than warranted.

The next sections analyze these effects and propose mitigating solutions. Due to space limitations, complete proofs are deferred to the Appendix.

6 Unfairness in Data Release

This section studies the effects of post-processing in a common data-release setting, where the goal is to release population counts that must also sum up to a public constant C. The section first introduces the projection mechanisms used to restore non-negativity and other aggregate data invariants and then studies its fairness effects.

Projections are common post-processing methods central to many data-release applications, including energy [Fioretto *et al.*, 2019], transportation [Fioretto *et al.*, 2018], and census data [Abowd *et al.*, 2019]. They are defined as:

$$\pi_{S+}(\tilde{\boldsymbol{x}}) := \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{K}_{S+}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2 , \qquad (P_{S+})$$

with feasible region defined as

$$\mathcal{K}_{S+} = \Big\{ \boldsymbol{v} \mid \sum_{i=1}^n v_i = C\,,\ \boldsymbol{v} \geqslant \mathbf{0} \Big\}.$$

Notice that P_{S+} is a convex program, and its unique optimal solution $\pi_{S+}(\tilde{\boldsymbol{x}})$ guarantees the desired data invariants by definition. For the analysis, it is also useful to consider a modified version P_S of P_{S+}, which differs from the latter only in that it ignores the non-negativity constraint $\boldsymbol{v} \geqslant \mathbf{0}$. Its feasible region and optimal solution are denoted, respectively, $\mathcal{K}_S$ and $\pi_S(\tilde{\boldsymbol{x}})$.

This section provides tight upper and lower bounds of the unfairness arising from projection operators. Lemma 3 and 4 are critical components to derive the α-fairness bounds developed in Theorem 1. The tightness of the proposed bounds is demonstrated in Example 1 and the existence of inherent unfairness in Example 2. Proposition 2 then presents an efficient method to evaluate the α-fairness bounds under the Gaussian mechanism, giving a uniquely valuable tool to decision makers to evaluate the impact of post-processing the data in their applications. To ease notation, the section omits the second argument P of the bias term $\mathcal{B}$ (as the P is an identity function for data-release settings). Additionally, unless otherwise specified, it assumes that the noisy data $\tilde{\boldsymbol{x}}$ is an output of the Laplace mechanism with parameter λ or the Gaussian mechanism with parameter σ.

Lemma 1. *[Zhu* et al.*, 2021] For any noisy data $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$, the closed-form solution $\pi_S(\tilde{\boldsymbol{x}})$ to program (P_S) is,*

$$\pi_S(\tilde{\boldsymbol{x}})_i = x_i + \eta_i - \frac{\sum_{j=1}^n \eta_j}{n} = \tilde{x}_i + \frac{C - \sum_{j=1}^n \tilde{x}_j}{n},$$

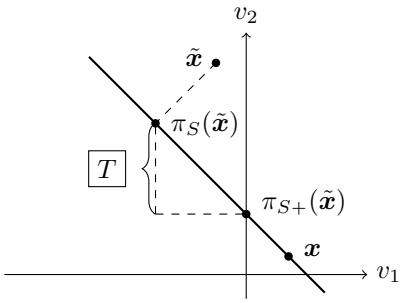

Figure 3: Illustration of different post-processed counts of $\tilde{\boldsymbol{x}}$. The solid line represents the feasible region $\mathcal{K}_S$ of program (P_S).

for any $i \in [n]$, with injected noise $\boldsymbol{\eta} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$.

Unlike $\pi_S(\tilde{\boldsymbol{x}})$, the post-processed count $\pi_{S+}(\tilde{\boldsymbol{x}})$ does not have a close-form expression. However, the following lemma introduces an implicit expression of $\pi_{S+}(\tilde{\boldsymbol{x}})$ on the basis of $\pi_S(\tilde{\boldsymbol{x}})$, establishing the foundation for the fairness analysis of the post-processing mechanism π_{S+}.

Lemma 2. *For any noisy data $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$, the solution $\pi_{S+}(\tilde{\boldsymbol{x}})$ to program $(P_{\mathrm{S}+})$ can be expressed as*

$$\pi_{S+}(\tilde{\boldsymbol{x}}) = \left(\pi_S(\tilde{\boldsymbol{x}}) - T(\pi_S(\tilde{\boldsymbol{x}})) \cdot \mathbf{1}\right)_{\geqslant 0} ,$$

where $(\cdot)_{\geqslant 0} = \max\{\cdot, 0\}$, and $T(\pi_S(\tilde{\boldsymbol{x}}))$ is the non-negative scalar that is the unique solution to the following equation

$$\sum_{i=1}^{n} \left(\pi_S(\tilde{\boldsymbol{x}})_i - T(\pi_S(\tilde{\boldsymbol{x}}))\right)_{\geqslant 0} = C .$$

Figure 3 provides an illustrative example relating the two post-processing mechanisms and the role of $T(\pi_S(\tilde{\boldsymbol{x}}))$. Given the noisy data $\tilde{\boldsymbol{x}}$ (top of the figure) π_S first projects it onto the solid line, which is the feasible region $\mathcal{K}_S$. Then, T needs to be deducted from both entries of $\pi_S(\tilde{\boldsymbol{x}})$ such that the positive part of $\pi_S(\tilde{\boldsymbol{x}}) - T \cdot \mathbf{1}$ equals $\pi_{S+}(\tilde{\boldsymbol{x}})$.

The following lemma provides lower and upper bounds for the bias difference of post-processing π_{S+} and plays a critical role in establishing the main results.

Lemma 3. *For any pair (i, j) such that $x_i \leqslant x_j$, the following relations hold,*

$$\mathcal{B}\left(\pi_{S+}\right)_i - \mathcal{B}\left(\pi_{S+}\right)_j \geqslant \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_i - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_j \tag{2a}$$

$$\mathcal{B}\left(\pi_{S+}\right)_i - \mathcal{B}\left(\pi_{S+}\right)_j \leqslant \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_i - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_j + \mathbb{E}_{\pi_S(\tilde{\boldsymbol{x}})}\left[T(\pi_S(\tilde{\boldsymbol{x}}))\right] \tag{2b}$$

with T defined as in Lemma 2 and $\mathcal{B}\left((\pi_S)_{\geqslant 0}\right)$ used as shorthand for $\mathbb{E}_{\tilde{\boldsymbol{x}}}\left[(\pi_S(\tilde{\boldsymbol{x}}))_{\geqslant 0}\right] - \boldsymbol{x}$.

While important, the upper bound (2b) is dependent on function T, which does not have a close-form expression; this makes it difficult to evaluate it. The following proposition provides an upper bound of T using $\pi_S(\tilde{\boldsymbol{x}})$.

Proposition 1. *For any noisy data $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$, $T(\pi_S(\tilde{\boldsymbol{x}}))$ is upper bounded by the sum of negative parts in $\pi_S(\tilde{\boldsymbol{x}})$:*

$$T(\pi_S(\tilde{\boldsymbol{x}})) \leqslant \sum_{i=1}^{n} \left(\pi_S(\tilde{\boldsymbol{x}})_i\right)_- ,$$

where $(\cdot)_- = -\min\{\cdot, 0\}$ takes the negative part of the input.

The following lemma presents an upper bound of the difference between biases: unlike the bound developed in Lemma 3, this new bound is independent of the injected noise.

Lemma 4. *For any pair (i, j) such that $x_i \leqslant x_j$, the following relation holds.*

$$\mathcal{B}\left(\pi_{S+}\right)_i - \mathcal{B}\left(\pi_{S+}\right)_j \leqslant x_j - x_i . \tag{3}$$

The next theorem is the main result of this section: it bounds the unfairness resulting from the projection mechanism π_{S+}. Without loss of generality, the true data $\boldsymbol{x}$ is assumed to be sorted in an increasing order, i.e., $x_i \leqslant x_j$, for any $i < j$.

Theorem 1. *The fairness bound α associated with the post-processed mechanism π_{S+} is bounded from the below by*

$$\alpha \geqslant \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_1 - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_n ,$$

and bounded from the above by

$$\alpha \leqslant \min\{x_n - x_1, \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_1 - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_n + \sum_{i=1}^{n} \mathbb{E}_{\pi_S(\tilde{\boldsymbol{x}})}\left[(\pi_S(\tilde{\boldsymbol{x}})_i)_-\right]\} .$$

Proof Sketch. By Equation (2a) in Lemma 3, notice that $\mathcal{B}\left(\pi_{S+}\right)_1$ is the largest entry while $\mathcal{B}\left(\pi_{S+}\right)_n$ is the smallest one among all the biases. The lower bound of the fairness bound α can then be derived in the following way.

$$\begin{aligned}\alpha &= \max_{j\in[n]} \mathcal{B}\left(\pi_{S+}\right)_j - \min_{j\in[n]} \mathcal{B}\left(\pi_{S+}\right)_j \\ &= \mathcal{B}\left(\pi_{S+}\right)_1 - \mathcal{B}\left(\pi_{S+}\right)_n \geqslant \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_1 - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_n .\end{aligned}$$

Likewise, Lemma 3 and 4, along with Proposition 1, make the joint effort to generate the upper bound. □

The tightness of the derived bounds follows from the following instance.

Example 1 (Centroid)**.** *The lower and upper bounds proposed in Theorem 1 hold with equality when the true data $\boldsymbol{x}$ is exactly the centroid of the feasible region $\mathcal{K}_{S+}$ of program $(P_{\mathrm{S}+})$, i.e., $\boldsymbol{x} = [C/n \ \ldots \ C/n] \in \mathbb{R}^n$, and the noisy data $\tilde{\boldsymbol{x}}$ is an output of either Laplace or Gaussian mechanism. In this case, the fairness bound α and its bounds in Theorem 1 happen to be 0, which also means that there is no fairness violation.*

The next example shows that post-processing definitely introduces unfairness when the true is not at the centroid.

Example 2 (Non-centroid)**.** *Suppose that the true data $\boldsymbol{x}$ is not the centroid of the feasible region $\mathcal{K}_{S+}$, i.e., $x_n > x_1$. The fairness bound α associated with the post-processing mechanism π_{S+} is strictly positive, i.e.,*

$$\alpha \geqslant \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_1 - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_n > 0 .$$

This negative result motivates the development of novel post-processing mechanisms in downstream decision processes, which are topics of the next section. The last result of this section provides an efficient evaluation of the proposed bounds via numerical integration methods.

Mechanism $\mathcal{M}$	α-fairness	Lower	Upper
Laplace	0.0245	0.0242	0.0288
Gaussian	0.0910	0.0897	0.1085

Table 1: Case study of Hawaii.

Proposition 2. *Let $\tilde{\boldsymbol{x}}$ be the output of the Gaussian mechanism with parameter σ. The key component of both lower and upper bounds in Theorem 1 can be written as*

$$\mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_1 - \mathcal{B}\left((\pi_S)_{\geqslant 0}\right)_n = \int_{-x_n}^{-x_1} \Phi(at)dt \in \left[\Phi\left(-ax_n\right)(x_n - x_1), \Phi\left(-ax_1\right)(x_n - x_1)\right],$$

where $a = \frac{1}{\sigma}\sqrt{\frac{n}{n-1}}$, and $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function.

It is interesting to demonstrate the tightness of these bounds using the US Census households counts at the county level for the state of Hawaii.

Example 3 (Hawaii)**.** *The state of Hawaii has a total number of $C = 453,558$ households distributed in $n = 5$ counties. The experiments use the Laplace mechanism with parameter $\lambda = 10$ and the Gaussian mechanism with parameter $\sigma = 25$. The empirical studies of α-fairness and its bounds in Theorem 1 associated with the post-processing mechanism π_{S+} over $1,000,000$ independent runs are reported in Table 1. The bounds of Gaussian mechanism use Proposition 2; those of Laplace mechanism are generated by the empirical means.*

The derived lower and upper bounds are really tight and provide decision makers a uniquely valuable tool to assess the unfairness introduced by post-processing.

7 Mechanisms for Downstream Decisions

Having shown that unfairness is unavoidable in common data-release settings, this section aims at designing post-processing mechanisms for decision processes that minimize their fairness impact on the resulting decisions. The mechanisms studied are tailored for the allocation problem P^F described in Section 5, which captures a wide class of resource allocation problems.

A natural baseline, currently adopted in census data-release tasks, is to first post-process the noisy data to meet the feasibility requirement (i.e., non-negativity) and then apply the allocation formula P^F to the post-processed counts. To restore feasibility, it suffices to take the positive part of $\tilde{\boldsymbol{x}}$ to obtain $(\tilde{\boldsymbol{x}})_{\geqslant 0}$, or equivalently, project $\tilde{\boldsymbol{x}}$ onto the non-negative orthant $\mathbb{R}^n_+$.

Definition 3 (Baseline Mechanism (BL))**.** *The* baseline mechanism *outputs, for each $i \in [n]$,*

$$\pi_{\mathrm{BL}}\left(\tilde{\boldsymbol{x}}\right)_i := \frac{a_i \cdot (\tilde{x}_i)_{\geqslant 0}}{\sum_{j=1}^n a_j \cdot (\tilde{x}_j)_{\geqslant 0}}.$$

It is possible to derive results similar to Example 2 for $(\cdot)_{\geqslant 0}$ when the baseline mechanism is used to produce feasible released data. Additionally, as shown in [Tran *et al.*, 2021], the disparate errors resulting from $(\cdot)_{\geqslant 0}$ can be further exacerbated when they are used as inputs to downstream decision problems. It suggests that the baseline mechanism might not be a good candidate for mitigating unfairness in this allocation problem. To address this limitation, consider the optimal post-processing mechanism in this context, i.e.,

$$\pi^* := \operatorname*{arg\,min}_{\pi \in \Pi_{\Delta_n}} \left\|\mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\pi(\tilde{\boldsymbol{x}}) - P^F\left(\boldsymbol{x}\right)\right]\right\|_{\rightleftharpoons}, \tag{4}$$

where $\Pi_{\Delta_n} = \{\pi \mid \pi : \mathbb{R}^n \mapsto \Delta_n\}$ represents a class of post-processing mechanisms whose images belong to the probability simplex Δ_n. The optimization problem in Equation (4) is intractable in its direct form, since $P^F\left(\boldsymbol{x}\right)$ is not available to the mechanism, motivating the need to approximate the objective function. Consider the following proxy $\mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\left\|\pi(\tilde{\boldsymbol{x}}) - P^F\left(\tilde{\boldsymbol{x}}\right)\right\|_{\rightleftharpoons}\right]$, which first exchanges the order of expectation and $\|\cdot\|_{\rightleftharpoons}$ and then replaces the true allocation $P^F\left(\boldsymbol{x}\right)$ with its noisy variant $P^F\left(\tilde{\boldsymbol{x}}\right)$. Then, the optimal post-processing mechanism π^*_α associated with this new proxy function becomes:

$$\pi^*_\alpha(\tilde{\boldsymbol{x}}) := \operatorname*{arg\,min}_{\boldsymbol{v} \in \Delta_n} \left\|\boldsymbol{v} - P^F\left(\tilde{\boldsymbol{x}}\right)\right\|_{\rightleftharpoons} \tag{P_α}$$

A mechanism, which is closely related to program (P_α), is presented as follows.

Definition 4 (Projection onto Simplex Mechanism (PoS))**.** *The* projection onto simplex mechanism *outputs the allocation as follows.*

$$\pi_{\mathrm{PoS}}\left(\tilde{\boldsymbol{x}}\right) := \operatorname*{arg\,min}_{\boldsymbol{v} \in \Delta^n} \left\|\boldsymbol{v} - P^F\left(\tilde{\boldsymbol{x}}\right)\right\|_2 \tag{P_{PoS}}$$

Program (P_{PoS}) projects $P^F\left(\tilde{\boldsymbol{x}}\right)$, which is not necessarily an allocation since it may violate non-negativity constraints, onto the closest feasible allocation. The next theorem establishes the equivalence between program (P_α) and program (P_{PoS}): It leads to a near-optimal post-processing mechanism. (The missing proofs of the rest of this paper can be found in the Appendix).

Theorem 2. *For any noisy data $\tilde{\boldsymbol{x}}$, the mechanism $\pi_{\mathrm{PoS}}\left(\tilde{\boldsymbol{x}}\right)$ generates the unique optimal solution to program (P_α).*

Figure 4 visualizes the resulting biases of the Title I allocation associated with these two mechanisms, π_{PoS} and π_{BL}. It is noteworthy that these two mechanisms achieve roughly same performance for the school districts that are allocated small amounts. However, under the baseline mechanism, the school districts that account for a significant portion of total budget receive much less funding than what they are supposed to receive when no differential privacy is applied. This is not the case for mechanism π_{PoS}, which reduces unfairness significantly. Recall that the notion of α-fairness measures the maximum difference among biases associated with different entities. Pictorially, the biases associated with π_{PoS} do not vary as drastically as the baseline mechanism. Table 2 quantifies the benefits of π_{PoS} over π_{BL}.

8 Generalizations

The results in Section 7 can be generalized to other fairness metrics. This section discusses an important metric that quantifies the extra budget needed to ensure that all of the problem

Privacy Budgets	$\epsilon = 0.1$		$\epsilon = 0.01$		$\epsilon = 0.001$	
Mechanisms	π_{BL}	π_{PoS}	π_{BL}	π_{PoS}	π_{BL}	π_{PoS}
α-fairness	3.00E-07	**1.50E-07**	1.70E-05	**1.75E-06**	8.06E-04	**2.23E-05**
Cost of Privacy	1.62E-05	**1.41E-05**	1.33E-03	**1.04E-03**	5.90E-02	**3.49E-02**

Table 2: Comparison between the two post-processing mechanisms in terms of two fairness metrics for different privacy budgets. This work takes Laplace mechanism and 200, 000 independent runs for numerical evaluation.

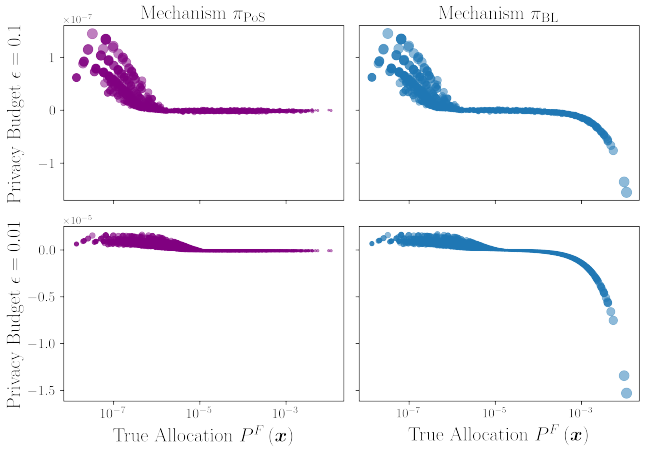

Figure 4: Illustration of the empirical biases (y-axis) associated with the two mechanisms π_{PoS} and π_{BL} (columns) for different privacy budgets (rows) versus the portions of education funds (x-axis) schools districts are guaranteed in the allocation with the true data. The Laplace mechanism is used for privacy protection and each experiment is repeated for 200,000 times.

entities receive the resources (e.g., amounts of funds) they are warranted by law.

Definition 5 (Cost of privacy [Tran *et al.*, 2021])**.** *Given the mechanism π, the total budget B to distribute and the true data $\boldsymbol{x}$, the* cost of privacy *is defined as*

$$B^+ := \sum_{j \in \mathcal{J}^-} |\, \mathcal{B}\left(\pi, P^F\right)_j \,| \cdot B\,,$$

with the index set $\mathcal{J}^- := \left\{ j \mid \mathcal{B}\left(\pi, P^F\right)_j < 0 \right\}$.

The next proposition establishes the equivalence between the cost of privacy and the ℓ_1 norm of the bias when the image of the mechanism π is restricted to be the probability simplex.

Proposition 3 (Cost of privacy as a ℓ_1-norm)**.** *Suppose that π is a post-processing mechanism, which belongs to the class Π_{Δ_n}. The cost of privacy is a multiplier of the ℓ_1-norm of its bias, i.e.,*

$$B^+ = \frac{B}{2} \cdot \left\| \mathcal{B}\left(\pi, P^F\right) \right\|_1 .$$

Since the optimal post-processing is again intractable in its direct form, i.e., it cannot be solved as an optimization problem, its objective can be replaced by the proxy $B/2 \cdot \mathbb{E}_{\tilde{\boldsymbol{x}}}\left[\left\|\pi(\tilde{\boldsymbol{x}}) - P^F\left(\tilde{\boldsymbol{x}}\right)\right\|_1\right]$. Then, the optimal post-processing mechanism π^*_{CoP} associated with this proxy function is given by

$$\pi^*_{\text{CoP}}(\tilde{\boldsymbol{x}}) := \underset{\boldsymbol{v} \in \Delta_n}{\arg\min} \frac{B}{2} \cdot \left\| \boldsymbol{v} - P^F\left(\tilde{\boldsymbol{x}}\right) \right\|_1 . \qquad (P_{\text{CoP}})$$

The next theorem depicts the connection between P_{CoP} and the two post-processing mechanisms proposed in Section 7.

Theorem 3. *For any noisy data $\tilde{\boldsymbol{x}}$, the mechanism $\pi_{\text{PoS}}\left(\tilde{\boldsymbol{x}}\right)$ generates an optimal solution to program (P_{CoP}). For any noisy data $\tilde{\boldsymbol{x}}$ such that $\sum_{j=1}^n a_j \cdot \tilde{x}_j > 0$, mechanism $\pi_{\text{BL}}\left(\tilde{\boldsymbol{x}}\right)$ generates an optimal solution to program (P_{CoP}) as well.*

This theorem demonstrates that mechanism π_{PoS} always produces an optimal solution to program (P_{CoP}) while the baseline mechanism achieves optimality with high probability. Table 2 shows that π_{PoS} may significantly outperform the baseline mechanism, providing substantial reductions in the cost of privacy.

9 Discussion and Conclusion

This paper was motivated by the recognition that the disparate error impacts of post-processing of differentially private outputs are poorly understood. Motivated by Census applications, it took a first step toward understanding how and why post-processing may produce disparate errors in data release and downstream allocation tasks. The paper showed that a popular class of post-processing mechanisms commonly adopted to restore invariants during the release of population statistics are inherently unfair. It proposed a tight bound on the unfairness and discussed an efficient method to evaluate the disparate impacts. Motivated by these negative results, the paper studied how post-processed data affects downstream decisions under a fairness lens and how to contrast such effects. In this context, the paper proposed to release the noisy, non-post-processed data, and post-processing the output of the downstream decisions instead. It focused on an important class of resource allocation problems used to allot funds or benefits and proposed a novel (approximately) optimal post-processing mechanism that is effective in mitigating unfairness under different fairness metrics. The analysis was complemented with numerical simulation on funds allocation based on private Census data showing up to an order magnitude improvements on different accuracy disparity metrics.

These results may have strong implications with respect to fairness in downstream decisions and should inform statistical agencies about the advantage of releasing private non-post-processed data, in favor of designing post-processing methods directly applicable in the downstream decision tasks of interest.

Acknowledgements

This research is partially supported by the National Science Foundation (NSF 2133169 and NSF 2133284) and by a Google Research Scholar Program award. The opinions expressed are solely those of the authors.

References

[Abowd and Schmutte, 2019] John M Abowd and Ian M Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

[Abowd *et al.*, 2019] John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. *US Census Bureau*, 2019.

[Abowd, 2018] John M Abowd. The US Census Bureau adopts differential privacy. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867. ACM, 2018.

[Cummings *et al.*, 2019] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Proceedings of the Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2019.

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

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[Ekstrand *et al.*, 2018] Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Proceedings of Conference on Fairness, Accountability and Transparency*, pages 35–47, 2018.

[Fioretto *et al.*, 2018] Ferdinando Fioretto, Chansoo Lee, and Pascal Van Hentenryck. Constrained-based differential privacy for private mobility. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1405–1413, 2018.

[Fioretto *et al.*, 2019] Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Differential privacy for power grid obfuscation. *IEEE Transactions on Smart Grid*, 11(2):1356–1366, 2019.

[Jagielski *et al.*, 2018] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. *arXiv preprint arXiv:1812.02696*, 2018.

[Pujol *et al.*, 2020] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

[Sonnenberg, 2016] W. Sonnenberg. Allocating grants for title i. *U.S. Department of Education, Institute for Education Science*, 2016.

[Tran *et al.*, 2021] Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 560–566. ijcai.org, 2021.

[Zhu *et al.*, 2021] Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. Bias and variance of post-processing in differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11177–11184, 2021.