Improving the Privacy Loss Under User-Level DP Composition for Fixed Estimation Error

V. Arvind Rameshwar
India Urban Data Exchange Program Unit,
Indian Institute of Science
Bengaluru, India
arvind.rameshwar@gmail.com

Anshoo Tandon
India Urban Data Exchange Program Unit,
Indian Institute of Science
Bengaluru, India
anshoo.tandon@gmail.com

ABSTRACT

This paper considers the private release of statistics of several disjoint subsets of a datasets, under user-level ϵ-differential privacy (DP). In particular, we consider the user-level differentially private release of sample means and variances of speed values in several grids in a city, in a potentially sequential manner. Traditional analysis of the privacy loss due to the sequential composition of queries necessitates a privacy loss degradation by a factor that equals the total number of grids. Our main contribution is an iterative, instance-dependent algorithm, based on clipping the number of user contributions, which seeks to reduce the overall privacy loss degradation under a canonical Laplace mechanism, while not increasing the worst estimation error among the different grids. We test the performance of our algorithm on synthetic datasets and demonstrate improvements in the privacy loss degradation factor via our algorithm. We also demonstrate improvements in the worst-case error using a simple extension of a pseudo-user creation-based mechanism. An important component of this analysis is our exact characterization of the sensitivities and the worst-case estimation errors of sample means and variances incurred by clipping user contributions in an arbitrary fashion, which we believe is of independent interest.

KEYWORDS

user-level differential privacy, minimax error, composition, traffic datasets

1 INTRODUCTION

Several landmark works have demonstrated that queries about seemingly benign functions of a dataset that is not publicly available can compromise the identities of the individuals in the dataset (see, e.g., [13, 16]). Examples of such reconstruction attacks for the specific setting of traffic datasets, which this paper concentrates on, can be found in [14, 18]. In this context, the framework of differential privacy (DP) was introduced in [4], which aims to preserve the privacy of users when each user contributes at most one sample, even in the presence of additional side information. More recent work [12] considered the setting where users could contribute more than one sample and formalized the framework of *user-level* DP, which requires the statistical indistinguishability of the output generated by a private mechanism, where potentially all of a user's contributions could be altered, from the output of the mechanism on the original dataset.

Now, in the traditional setting of (pure) ϵ-DP or ϵ-user-level DP (where ϵ captures the privacy loss), the Basic Composition Theorem shows that if the user were to pose multiple queries to the data curator in a potentially sequential (or adaptive) manner, the total privacy loss degrades by a factor that, in the worst case, equals the number of queries (see [5, Cor. 3.15]). It is also well-known that there exists a differentially private mechanism, namely, the canonical Laplace mechanism, which achieves this privacy loss (see, e.g., [15, Sec. 2]). We mention that in the setting where we allow for (approximate) (ϵ, δ)-DP, for a certain range of parameter values, it is possible to obtain improvements in the worst-case privacy loss as compared to that guaranteed by basic composition [5, Sec. 3.5], [6, 10].

This paper differs in its results from other papers on composition in two respects: firstly, we consider the framework of user-level privacy; secondly, we work with pure ϵ-user-level-DP and provide an algorithm that seeks to reduce the worst-case privacy loss degradation, in an instance-dependent manner, while maintaining the worst-case estimation error. Our treatment, hence, is a study of composition of user-level DP mechanisms that *jointly* considers the errors due to noise addition for privacy and due to the *bias* that results from the estimator used in the DP mechanism being different from the true function to be released. Our focus on user-level privacy assumes significance in the context of most real-world IoT datasets, such as traffic databases, which record multiple contributions from every user, with different users contributing potentially different number of samples. We mention that recent work [2] presented some algorithms for real-world datasets, based on the work in [12] and [9], which guarantee user-level ϵ-DP, and also provided theoretical proofs of their performance trends.

Our main contribution in this work is the development of a simple, yet novel, iterative algorithm, for improving the overall privacy loss under composition of several user-level DP mechanisms, each of which releases the sample mean and variance of speed records in a particular grid in a city. Our algorithm achieves the claimed improvement in privacy loss by suppressing the contributions of selected users in selected grids, while not increasing the *largest worst-case* error across all the grids. Crucial components of the design of our algorithm are exact characterizations of the sensitivity of the sample variance and the worst-case errors (over all datasets) in the estimation of the sample mean and variance of each grid, when the contributions of some users have been suppressed. Furthermore, the exact user-level sensitivity of the sample variance

YYYY(X), 1–15

computed in this work yields, as a corollary, a strict improvement over the bound on *item-level* sensitivity in [4, p. 10], which is often taken as the standard in the DP literature. With the aid of our characterizations of the worst-case errors, we suggest a simple psuedo-user creation-based algorithm—a natural extension of the work in [2]—which helps reduce the worst-case estimation error. We emphasize that our algorithm can be applied more generally to the release of other statistics (potentially different from the sample mean and variance) of several disjoint subsets of the records in a dataset.

The paper is organized as follows: Section 2 presents the problem formulation and recapitulates preliminaries on DP and user-level DP. Section 3 contains a description of the mechanisms of importance to this paper and presents an exact characterization of the (user-level) sensitivity of the sample variance function. Section 4 exactly characterizes the worst-case errors in the estimation of sample mean and variance due to the suppression of selected records. Section 5 then describes our main algorithm that suppresses user contributions in an effort to improve the privacy loss under composition. We then numerically evaluate the performance of our algorithm on synthetically generated datasets in terms of the privacy loss degradation, in Section 6, and suggest a simple pseudo-user creation-based algorithm to improve the worst-case estimation error, over all grids. The paper is concluded in Section 7 with some directions for future research.

2 PRELIMINARIES

2.1 Notation

For a given $n \in \mathbb{N}$, the notation $[n]$ denotes the set $\{1, 2, \ldots, n\}$ and the notation $[a : b]$ denotes the set $\{a, a+1, \ldots, b\}$, for $a, b \in \mathbb{N}$ and $a \leq b$. Given a length-n vector $\mathbf{u} \in \mathbb{R}^n$, we define $\|\mathbf{u}\|_1 := \sum_{i=1}^{n} |u_i|$ to be the ℓ_1-norm of the vector $\mathbf{u}$. We write $X \sim P$ to denote that the random variable X is drawn from the distribution P. We use the notation $\text{Lap}(b)$ to refer to a random variable X drawn from the zero-mean Laplace distribution with standard deviation $\sqrt{2}b$; its probability distribution function (p.d.f.) obeys

$$f_X(x) = \frac{1}{2b} e^{-|x|/b},\ x \in \mathbb{R}.$$

2.2 Problem Setup

This work is motivated by the analysis of traffic datasets, which contain records of the data provided by IoT sensors deployed in a city, pertaining to information on vehicle movement. Each record catalogues, typically among other information, the license plate of the vehicle, the location at which the data was recorded, a timestamp, and the actual data value itself, which is the speed of the bus. Most data analysis tasks on such datasets proceed as follows: first, in an attempt to obtain fine-grained information about the statistics of the speed samples in different areas of the city, the total area of the city is divided into hexagon-shaped grids (see, e.g., Uber's Hexagonal Hierarchical Spatial Indexing System [1], which provides an open-source library for such partitioning tasks). Next, the timestamps present in the data records are quantized (or binned) into timeslots of fixed duration (say, one hour). In this work, we seek to release the sample averages and sample variances of speeds of vehicles in all the grids that the city area has been divided into, privately (and potentially adaptively), to a client who has no prior knowledge of these values. We remark that the algorithms discussed in this paper are readily applicable to general spatio-temporal IoT datasets with bounded data samples, for releasing other differentially private statistics.

2.3 Problem Formulation

Let $\mathcal{L}$ denote the collection of all users (or equivalently, distinct license plates) in the city, and let $\mathcal{G}$ be the collection of grids that the city area has been divided into. We set $L := |\mathcal{L}|$ and $G := |\mathcal{G}|$. Furthermore, for each user $\ell \in \mathcal{L}$ and each grid $g \in \mathcal{G}$, we let ${}^{g}m_\ell$ denote the (non-negative integer) number of speed samples contributed by user ℓ in the records corresponding to grid g. Now, for a given user $\ell \in \mathcal{L}$, let $m_\ell := \sum_{g\in\mathcal{G}} {}^{g}m_\ell$ be the total number of speed samples contributed by user ℓ across all grids. Next, for every grid $g \in \mathcal{G}$, let

$${}^{g}\mathcal{L} := \{\ell \in \mathcal{L} : {}^{g}m_\ell > 0\}$$

be the collection of users whose contributions constitute the data records corresponding to grid g. We let ${}^{g}m^\star$ (resp. ${}^{g}m_\star$) denote the largest (resp. smallest) number of samples contributed by any user in grid $g \in \mathcal{G}$. Formally, ${}^{g}m^\star = \max_{\ell\in{}^{g}\mathcal{L}} {}^{g}m_\ell$, and ${}^{g}m_\star = \min_{\ell\in{}^{g}\mathcal{L}} {}^{g}m_\ell$. For every user $\ell \in \mathcal{L}$, let

$$\mathcal{G}_\ell := \{g \in \mathcal{G} : {}^{g}m_\ell > 0\}$$

be the collection of grids whose records user ℓ contributes to. In line with the previous notation, we set ${}^{g}L := |{}^{g}\mathcal{L}|$ and $G_\ell := |\mathcal{G}_\ell|$. Throughout this paper, we assume, without loss of generality, that $G_1 \geq G_2 \geq \ldots \geq G_L$.

Now, let ${}^{g}S_\ell$ denote the vector of speed samples contributed by user $\ell \in \mathcal{L}$ in grid $g \in \mathcal{G}$; more precisely, ${}^{g}S_\ell := \left({}^{g}S_\ell^{(j)} : j \in [{}^{g}m_\ell]\right)$. We assume that each ${}^{g}S_\ell^{(j)}$ is a non-negative real number that lies in the interval $[0, U]$, where U is a fixed upper bound on the speeds of the vehicles. For the real-world datasets that are the objects of consideration in this paper, the speed samples are drawn according to some unknown distribution P that is potentially non-i.i.d. (independent and identically distributed) across samples and users. Our analysis is distribution-free in that we work with the worst-case errors in estimation over all datasets, in place of distribution-dependent error metrics such as the expected error (see, e.g., [11, Sec. 1.1] for a discussion).

We call the dataset consisting of the speed records contributed by users as

$$\mathcal{D} = \left\{\left(\ell, \left\{{}^{g}S_\ell : g \in \mathcal{G}\right\}\right) : \ell \in \mathcal{L}\right\}.$$

We let D denote the universe of all possible datasets with a given distribution of numbers of samples contributed by users across grids $\{{}^{g}m_\ell : \ell \in \mathcal{L}, g \in \mathcal{G}\}$.

The function that we are interested in is the length-G vector $f : \mathsf{D} \to (\mathbb{R}^2)^G$, each of whose components is a 2-tuple of the sample average and the sample variance of speed samples in each grid. More precisely, we have

$$f(\mathcal{D}) = \left({}^{g}f(\mathcal{D}) : g \in \mathcal{G}\right), \tag{1}$$

where ${}^{g}f : \mathsf{D} \to \mathbb{R}^2$ is such that

$${}^{g}f(\mathcal{D}) = \begin{bmatrix} {}^{g}\mu(\mathcal{D}) \\ {}^{g}\mathsf{Var}(\mathcal{D}) \end{bmatrix}. \tag{2}$$

Here,

$$ {}^g\mu(\mathcal{D}) := \frac{1}{\sum\limits_{\ell \in {}^g\mathcal{L}} {}^g m_\ell} \cdot \sum_{\ell \in {}^g\mathcal{L}} \sum_{j=1}^{{}^g m_\ell} {}^g S_\ell^{(j)} \tag{3} $$

is the sample mean of speed records in grid g and

$$ {}^g\mathrm{Var}(\mathcal{D}) = \frac{1}{\sum_{\ell \in {}^g\mathcal{L}} {}^g m_\ell} \cdot \sum_{\ell \in {}^g\mathcal{L}} \sum_{j=1}^{{}^g m_\ell} \left({}^g S_\ell^{(j)} - {}^g\mu(\mathcal{D}) \right)^2 \tag{4} $$

is the sample variance of speed records in grid g. For the purposes of this work, one can equivalently think of $f(\mathcal{D})$ as a length-$2G$ vector, each of whose components is a scalar mean or variance. A central objective in user-level differential privacy is the private release of an estimate of f, without compromising too much on the accuracy in estimation. We next recapitulate the definition of user-level differential privacy [12].

2.4 User-Level Differential Privacy

Consider two datasets $\mathcal{D}_1 = \{(\ell, \{{}^g x_\ell : g \in \mathcal{G}\}) : \ell \in \mathcal{L}\}$ and $\mathcal{D}_2 = \{(\ell, \{{}^g \tilde{x}_\ell : g \in \mathcal{G}\}) : \ell \in \mathcal{L}\}$ consisting of the same users, with each user contributing the same number of (potentially different) data values. Recall that D is the universal set of such databases. We say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are "user-level neighbours" if there exists $\ell_0 \in [L]$ such that $\left\{{}^g x_{\ell_0} : g \in \mathcal{G}\right\} \neq \left\{{}^g \tilde{x}_{\ell_0} : g \in \mathcal{G}\right\}$, with $\{{}^g x_\ell : g \in \mathcal{G}\} = \{{}^g \tilde{x}_\ell : g \in \mathcal{G}\}$, for all $\ell \neq \ell_0$. Clearly, datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ differ in at most m_{ℓ_0} samples, where $m_{\ell_0} \le m^\star$, where $m^\star := \max_{\ell \in \mathcal{L}} m_\ell$.

Definition 2.1. For a fixed $\varepsilon > 0$, a mechanism $M : \mathrm{D} \to \mathbb{R}^d$ is said to be user-level ε-DP if for every pair of datasets $\mathcal{D}_1, \mathcal{D}_2$ that are user-level neighbours, and for every measurable subset $Y \subseteq \mathbb{R}^d$, we have that

$$ \Pr[M(\mathcal{D}_1) \in Y] \le e^\varepsilon \Pr[M(\mathcal{D}_2) \in Y]. $$

Next, we recall the definition of the user-level sensitivity of a function of interest.

Definition 2.2. Given a function $\theta : \mathrm{D} \to \mathbb{R}^d$, we define its user-level sensitivity Δ_θ as

$$ \Delta_\theta := \max_{\mathcal{D}_1, \mathcal{D}_2 \text{ u-l nbrs.}} \|\theta(\mathcal{D}_1) - \theta(\mathcal{D}_2)\|_1 , $$

where the maximization is over datasets that are user-level neighbours.

In this paper, we use the terms "sensitivity" and "user-level sensitivity" interchangeably. The next result is well-known and follows from standard DP results [4, Prop. 1][1]:

Theorem 2.1. *For a function $\theta : \mathrm{D} \to \mathbb{R}^d$, the mechanism $M^{Lap} : \mathrm{D} \to \mathbb{R}^d$ defined by*

$$ M^{Lap}(\mathcal{D}) = \theta(\mathcal{D}) + Z, $$

where $Z = (Z_1, \ldots, Z_d)$ is such that $Z_i \sim Lap(\Delta_\theta/\varepsilon)$, is user-level ε-DP.

Furthermore, by standard results on the tail probabilities of Laplace random variables, we obtain the following bound on the estimation error due to the addition of noise for privacy:

Proposition 2.1. *For a given function $\theta : \mathrm{D} \to \mathbb{R}^d$ and for any dataset $\mathcal{D}_1$, we have that*

$$ \Pr\left[\left\| M^{Lap}(\mathcal{D}_1) - \theta(\mathcal{D}_1) \right\|_1 \ge \frac{\Delta_\theta \ln(1/\delta)}{\varepsilon} \right] \le \delta, $$

for all $\delta \in (0, 1]$.

In the following subsection, we shall discuss the overall privacy loss that results from the composition of several user-level ϵ-DP mechanisms together.

2.5 Composition of User-Level DP Mechanisms

Recall that our objective in this work is the (potentially sequential, or adaptive) release of a fixed function (in particular, the sample mean and sample variance) of the records in each grid, over all grids. The following fundamental theorem from the DP literature [5, Cor. 3.15] captures the worst-case privacy loss degradation upon composition of (user-level) DP mechanisms. For each $g \in \mathcal{G}$, let ${}^g M : \mathrm{D} \to \mathbb{R}^d$ be an ${}^g\epsilon$-DP algorithm that acts exclusively on those records from grid g. Further, let $M = ({}^g M : g \in \mathcal{G})$ be the composition of the G mechanisms above.

Theorem 2.2 (Basic Composition Theorem). *We have that M is user-level $\sum_{g \in \mathcal{G}} {}^g\epsilon$-DP.*

It is well-known (see, e.g., [15, Sec. 2.1]) that Theorem 2.2 is tight, in that there exists a *Laplace mechanism* (of the form in Theorem 2.1) that achieves a privacy loss of $\sum_{g \in \mathcal{G}} {}^g\epsilon$ upon composition.

Observe from Theorem 2.2 that in the case when ${}^g\epsilon = \epsilon$, for all $g \in \mathcal{G}$, we obtain an overall privacy loss of $G\epsilon$, upon composition. Clearly, when the number of grids G is large, the overall privacy loss is large, as well.

We next present a simple improvement of the Basic Composition Theorem above that takes into account the fact that each mechanism ${}^g M$, for $g \in \mathcal{G}$, acts only on the records in the grid g. Let $\overline{\epsilon} := \max_{\ell \in \mathcal{L}} \sum_{g \in \mathcal{G}_\ell} {}^g\epsilon$.

Theorem 2.3. *We have that M is user-level $\overline{\epsilon}$-DP.*

Proof. Consider datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ (exclusively) in the contributions of user $\ell \in \mathcal{L}$. Now, consider any measurable set $T = \left({}^1T, \ldots, {}^GT\right) \subseteq \mathbb{R}^G$. For ease of reading, we let ${}^{(g-1)}M(\mathcal{D}) := ({}^1M(\mathcal{D}), \ldots, {}^{g-1}M(\mathcal{D})))$; likewise, we let ${}^{(g-1)}T := ({}^1T, \ldots, {}^{g-1}T)$.

$$ \begin{aligned} &\frac{\Pr[M(\mathcal{D}) \in T]}{\Pr[M(\mathcal{D}') \in T]} \\ &= \frac{\prod_{g \in \mathcal{G}} \Pr[{}^g M(\mathcal{D}) \in {}^g T | {}^{(g-1)}M(\mathcal{D}) \in {}^{(g-1)}T]}{\prod_{g \in \mathcal{G}} \Pr[{}^g M(\mathcal{D}') \in {}^g T | {}^{(g-1)}M(\mathcal{D}') \in {}^{(g-1)}T]} \\ &\le e^{\sum_{g \in \mathcal{G}_\ell} {}^g\epsilon}, \end{aligned} $$

where the last inequality follows from the DP property of each mechanism ${}^g M$, $g \in \mathcal{G}$. The result then follows immediately. □

As a simple corollary, from our assumption that $G_1 \ge G_2 \ge \ldots \ge G_L$, we obtain the following result:

[1] It is well-known that it is sufficient to focus on noise-adding DP mechanisms. The assumption that our mechanisms are *additive-noise* or *noise-adding* mechanisms is without loss of generality, since it is known that every privacy-preserving mechanism can be thought of as a noise-adding mechanism (see [7, Footnote 1] and [8]). Moreover, under some regularity conditions, for small ϵ (or equivalently, high privacy requirements), it is known that Laplace distributed noise is asymptotically optimal in terms of the magnitude of error in estimation [7, 8].

Corollary 2.1. *When ${}^g\epsilon = \epsilon$, for all $g \in \mathcal{G}$, we have that M is $G_1\epsilon$-DP.*

In what follows, we shall focus on this simplified setting where the privacy loss ${}^g\epsilon$ for each grid g is fixed to be $\epsilon > 0$. Note that if G_1 is large, the privacy loss upon composing the mechanisms corresponding to the different grids of the city is correspondingly large.

A natural question that arises, hence is: can we improve the worst-case privacy loss (in the sense of Corollary 2.1) in such a manner as to preserve some natural notion of the worst-case error over all grids? In what follows, we shall show that for a specific class of (canonical) mechanisms, a notion of the worst-case error over all grids can be made precise, which will then serve as a guideline for the design of our algorithm that improves the privacy loss degradation by clipping user contributions.

We end this subsection with a remark. In the setting of *item-level* DP, where each user contributes at most one sample, it follows from Theorem 2.1 that the composition of mechanisms that act on *disjoint* subsets of a dataset has the same privacy loss as that of any individual mechanism, i.e., M is ϵ-DP as well. In such a setting, it is not possible to improve on the privacy loss degradation by clipping user contributions.

The next section describes the mechanisms that will be of use in this paper; we refer the reader to [2, 12] for more user-level DP mechanisms for releasing sample means and their performance on real-world datasets.

3 MECHANISMS FOR RELEASING DP ESTIMATES

In this section and the next, we focus our attention on a single grid $g \in \mathcal{G}$. For notational simplicity, we shall drop the explicit dependence of the notation (via superscripts) in Section 2 on g; alternatively, it is instructive to consider this setting as a special case of the setting in Section 2, where $|\mathcal{G}| = 1$. In particular, ${}^g m_\ell = m_\ell$, for all $\ell \in \mathcal{L}$, ${}^g\mathcal{L} = \mathcal{L}$, and ${}^g\mu = \mu$ and ${}^g\mathsf{Var} = \mathsf{Var}$. With some abuse of notation, we let $\mathcal{D}$ denote the dataset consisting of records in grid g and let D denote the universal set of datasets with the distribution $\{m_\ell\}$ of user contributions.

We now describe two mechanisms for releasing user-level differentially private estimates of the sample mean and variance of a single grid.

3.1 Baseline

Given the definitions μ and Var as in (3) and (4), the first mechanism, which we call Baseline, simply adds the right amount of Laplace noise to μ and Var to ensure user-level ϵ-DP. Formally, the Baseline mechanism $M_\mathrm{b} : \mathsf{D} \to \mathbb{R}^2$ obeys

$$M_\mathrm{b}(\mathcal{D}) = \begin{bmatrix} M_{\mu,\mathrm{b}}(\mathcal{D}) \\ M_{\mathsf{Var},\mathrm{b}}(\mathcal{D}) \end{bmatrix},$$

where

$$M_{\mu,\mathrm{b}}(\mathcal{D}) = \mu(\mathcal{D}) + \mathrm{Lap}(2\Delta_\mu/\varepsilon),$$

and

$$M_{\mathsf{Var},\mathrm{b}}(\mathcal{D}) = \mathsf{Var}(\mathcal{D}) + \mathrm{Lap}(2\Delta_{\mathsf{Var}}/\varepsilon).$$

Note that the privacy budget for the release of each of the sample mean and variance is fixed to $\epsilon/2$, leading to M_b being ϵ-user-level DP, overall, by Theorem 2.2. Furthermore, from the definition of user-level sensitivity in Section 2, we have that

$$\Delta_\mu = \frac{U \cdot \max_{\ell\in\mathcal{L}} m_\ell}{\sum\limits_{\ell\in\mathcal{L}} m_\ell} = \frac{U \cdot m^\star}{\sum\limits_{\ell\in\mathcal{L}} m_\ell}. \tag{5}$$

An explicit computation of the user-level sensitivity of Var, however, requires significantly more effort. The next proposition, whose proof is provided in Appendix A exactly identifies Δ_{Var}.

Proposition 3.1. *We have that*

$$\Delta_{Var} = \begin{cases} \frac{U^2\, m^\star(\sum_\ell m_\ell - m^\star)}{(\sum_\ell m_\ell)^2}, \ \text{if } \sum_\ell m_\ell > 2m^\star, \\ \frac{U^2}{4}, \ \text{if } \sum_\ell m_\ell \le 2m^\star \text{ and } \sum_\ell m_\ell \text{ is even,} \\ \frac{U^2}{4}\cdot\left(1 - \frac{1}{(\sum_\ell m_\ell)^2}\right), \ \text{if } \sum_\ell m_\ell \le 2m^\star \text{ and } \sum_\ell m_\ell \text{ is odd.} \end{cases}.$$

We then obtain the following corollary on the sensitivity of the sample variance function in the item-level DP setting where each user contributes exactly one sample, i.e., when $m_\ell = 1$, for all $\ell \in \mathcal{L}$.

Corollary 3.1. *In the setting of item-level DP, we have*

$$\Delta_{Var} = \begin{cases} \frac{U^2(L-1)}{L^2}, \ \text{if } L \ge 2, \\ 0, \ \text{if } L = 1. \end{cases}.$$

On the other hand, the well-known upper bound on the sensitivity of the sample variance in [4, p. 10] that is now standard for DP applications shows that in the item-level DP setting, $\Delta_{\mathsf{Var}} \le 8U^2/L$. Clearly, the exact sensitivity computed in Corollary 3.1 is a *strict* improvement over this bound, by a multiplicative factor of more than 8, for all L.

Now, consider the expression in Proposition 3.1 above, for a *fixed* $\sum_\ell m_\ell$. Suppose also that $\sum_\ell m_\ell > 2m^\star$. Hence, for this range of $m^\star$ values, it is easy to argue that $h(m^\star) := m^\star(\sum_\ell m_\ell - m^\star)$ is increasing in $m^\star$, implying that for a fixed value of $\sum_\ell m_\ell$, we have that Δ_{Var} is increasing in $m^\star$, in the regime where $\sum_\ell m_\ell > 2m^\star$. Furthermore, it is easy to argue that $\Delta_{\mathsf{Var}} \le U^2/4$, for all values of $\{m_\ell\}$, implying that Δ_{Var} is non-decreasing, overall, as $m^\star$ increases. In other words, a large value of $m^\star$ leads to a large sensitivity. In our next mechanism, which we call Clip, we attempt to ameliorate this issue by clipping the number of contributions of each user in the grid, at the cost of some error in accuracy.

3.2 Clip

We proceed to describe a simple modification of the previous mechanism, which we call Clip, for releasing user-level differentially private estimates of μ and Var, by clipping (or suppressing) selected records. For $\ell \in \mathcal{L}$, we let $\Gamma_\ell \in [0 : m_\ell]$ denote the number of contributions of user ℓ that have *not* been clipped; without loss of generality, we assume that the set of indices of these samples is $[\Gamma_\ell]$. Further, we assume that $\sum_\ell \Gamma_\ell > 0$. We use the notation $\Gamma^\star := \max_{\ell\in\mathcal{L}} \Gamma_\ell$.

Given the dataset $\mathcal{D}$, we set

$$\mu_{\mathrm{clip}}(\mathcal{D}) = \frac{1}{\sum_\ell \Gamma_\ell} \cdot \sum_{\ell=1}^{L} \sum_{j=1}^{\Gamma_\ell} S_\ell^{(j)} \tag{6}$$

to be that estimator of the sample mean that is obtained by retaining only Γ_ℓ samples, for each user ℓ. Next, we set

$$\mathsf{Var}_{\text{clip}}(\mathcal{D}) = \frac{1}{\sum_\ell \Gamma_\ell} \cdot \sum_{\ell=1}^{L} \sum_{j=1}^{\Gamma_\ell} \left(S_\ell^{(j)} - \mu_{\text{clip}}(\mathcal{D})\right)^2 \tag{7}$$

to be an estimator of the sample variance that makes use of the previously computed estimator $\mu_{\text{clip}}(\mathcal{D})$ of the sample mean.

Our mechanism $M_{\text{clip}} : \mathsf{D} \to \mathbb{R}^2$ obeys

$$M_{\text{clip}}(\mathcal{D}) = \begin{bmatrix} M_{\mu,\text{clip}}(\mathcal{D}) \\ M_{\mathsf{Var},\text{clip}}(\mathcal{D}) \end{bmatrix}, \tag{8}$$

where

$$M_{\mu,\text{clip}}(\mathcal{D}) = \mu_{\text{clip}}(\mathcal{D}) + \text{Lap}(2\Delta_{\mu_{\text{clip}}}/\varepsilon),$$

and

$$M_{\mathsf{Var},\text{clip}}(\mathcal{D}) = \mathsf{Var}_{\text{clip}}(\mathcal{D}) + \text{Lap}(2\Delta_{\mathsf{Var}_{\text{clip}}}/\varepsilon).$$

Here, $\Delta_{\mu_{\text{clip}}}$ and $\Delta_{\mathsf{Var}_{\text{clip}}}$ are respectively the user-level sensitivities of the clipped mean estimator μ_{clip} and the clipped variance estimator $\mathsf{Var}_{\text{clip}}$. As before, we assign a privacy budget of $\epsilon/2$ for each of the mechanisms $M_{\mu,\text{clip}}$ and $M_{\mathsf{Var},\text{clip}}$. Clearly, both these algorithms are $\varepsilon/2$-user-level DP, from Theorem 2.1, resulting in the overall mechanism M_{clip} being ϵ-user-level DP, from Theorem 2.2.

By arguments similar to those in [2, Sec. III.C], we have that

$$\Delta_{\mu_{\text{clip}}} = \frac{U\,\Gamma^\star}{\sum_{\ell=1}^{L} \Gamma_\ell}. \tag{9}$$

Furthermore, by analysis entirely analogous to the proof of Proposition 3.1, we obtain the following lemma:

Lemma 3.1. *We have that*

$$\Delta_{Var_{clip}} = \begin{cases} \frac{U^2\ \Gamma_\ell^\star(\sum_\ell \Gamma_\ell - \Gamma_\ell^\star)}{(\sum_\ell \Gamma_\ell)^2}, & \textit{if } \sum_\ell \Gamma_\ell > 2\Gamma^\star, \\ \frac{U^2}{4}, & \textit{if } \sum_\ell \Gamma_\ell \le 2\Gamma^\star \textit{ and } \sum_\ell \Gamma_\ell \textit{ is even}, \\ \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell \Gamma_\ell)^2}\right), & \textit{if } \sum_\ell \Gamma_\ell \le 2\Gamma^\star \textit{ and } \sum_\ell \Gamma_\ell \textit{ is odd.} \end{cases}$$

In Appendix B, we show that for a special class of clipping strategies considered in [2], the sensitivities $\Delta_{\mu_{\text{clip}}}$ and $\Delta_{\mathsf{Var}_{\text{clip}}}$ are in fact at most the values of their Baseline counterparts Δ_μ and Δ_{Var}, respectively. We mention that the mechanisms $M_{\mu,\text{clip}}$ and $M_{\mathsf{Var},\text{clip}}$ are also called as pseudo-user creation-based mechanisms, in this case.

In the next section, we focus more closely on the Clip mechanism and explicitly characterize the worst-case errors (over all datasets) due to clipping the contributions of users.

4 WORST-CASE ERRORS IN ESTIMATION OF SAMPLE MEAN AND VARIANCE

In this section, we continue to focus on a single grid $g \in \mathcal{G}$ and explicitly identify the *worst-case* error due to clipping incurred, over all datasets, by the Clip mechanism with an arbitrary choice $\Gamma_\ell \in [0 : m_\ell]$, for $\ell \in \mathcal{L}$. The characterizations of worst-case errors in this section will be of use in the design of our algorithm for improving the privacy loss degradation under composition, via the clipping (or suppression) of user contributions in selected grids. We now make the notion of the worst-case clipping error formal.

Consider the functions μ, Var that stand for the true sample mean and variance, and the functions $\mu_{\text{clip}}, \mathsf{Var}_{\text{clip}}$ that stand for the sample mean and variance of the clipped samples, for some fixed values $\Gamma_\ell \in [0 : m_\ell]$, where $\ell \in \mathcal{L}$. We now define

$$E_\mu(\mathcal{D}) := |\mu(\mathcal{D}) - \mu_{\text{clip}}(\mathcal{D})|,$$

as the clipping error (or bias) for the mean on dataset $\mathcal{D}$, and

$$E_\mu := \max_{\mathcal{D} \in \mathsf{D}} E_\mu(\mathcal{D})$$

as the worst-case clipping error for the mean. Likewise, we define

$$E_{\mathsf{Var}}(\mathcal{D}) := |\mathsf{Var}(\mathcal{D}) - \mathsf{Var}_{\text{clip}}(\mathcal{D})|,$$

as the clipping error for the variance on dataset $\mathcal{D}$, and

$$E_{\mathsf{Var}} := \max_{\mathcal{D} \in \mathsf{D}} E_{\mathsf{Var}}(\mathcal{D})$$

as the worst-case clipping error for the variance. The following theorem from [2] then holds[2]:

Theorem 4.1 (Lemma V.1 in [2]). *We have that*

$$E_\mu = U \cdot \left(1 - \frac{\sum_\ell \Gamma_\ell}{\sum_\ell m_\ell}\right).$$

In what follows, we characterize exactly the worst-case error for the variance.

Theorem 4.2. *We have that $E_{Var} = 0$ if $\Gamma_\ell = m_\ell$, for all $\ell \in \mathcal{L}$. Furthermore, if $\sum_\ell \Gamma_\ell < \sum_\ell m_\ell$, we have*

$$E_{Var} = \begin{cases} \frac{U^2 \cdot \sum_\ell \Gamma_\ell \cdot \sum_{\ell'} (m_{\ell'} - \Gamma_{\ell'})}{(\sum_\ell m_\ell)^2}, & \textit{if } \sum_\ell m_\ell > 2\sum_\ell \Gamma_\ell, \\ \frac{U^2}{4}, & \textit{if } \sum_\ell m_\ell \le 2\sum_\ell \Gamma_\ell \textit{ and } \sum_\ell m_\ell \textit{ is even}, \\ \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell m_\ell)^2}\right), & \textit{if } \sum_\ell m_\ell \le 2\sum_\ell \Gamma_\ell \textit{ and } \sum_\ell m_\ell \textit{ is odd.} \end{cases}$$

The proof of Theorem 4.2 is provided in Appendix C.

5 AN ERROR METRIC AND AN ALGORITHM FOR CONTROLLING PRIVACY LOSS

In this section, we return to our original problem of releasing the sample means and variances of different grids in the city, possibly sequentially. We present our algorithm that seeks to control the privacy loss of a certain user-level DP mechanism for jointly releasing the sample mean and variance of all grids in the city, by clipping user contributions. As we shall see, the individual mechanisms for each grid simply add a suitable amount of Laplace noise that is tailored to the sensitivity of the functions in the grid *post* clipping. Our algorithm hence crucially relies on the analyses of the sensitivity and the worst-case clipping error of the Clip mechanism in Sections 3.2 and 4.

5.1 An Error Metric for Worst-Case Performance

We shall first formally define a notion of the worst-case error of any mechanism $M = ({}^gM : g \in \mathcal{G})$, over all datasets, and over all grids. Our algorithm will then follow naturally from these definitions.

[2] While [2] contained a proof of 4.1 for the special case when $\Gamma_\ell = \min\{m, m_\ell\}$, for $\ell \in \mathcal{L}$ and for some fixed $m \in [m_\star, m^\star]$, this theorem holds for general values $\Gamma_\ell \in [0 : m_\ell]$ as well.

Formally, consider a mechanism ${}^{g}M_\theta : \mathsf{D} \to \mathbb{R}^d$, for $g \in \mathcal{G}$, for the user-level differentially private release of a statistic ${}^{g}\theta : \mathsf{D} \to \mathbb{R}^d$ of the records in grid g. Suppose that ${}^{g}M_\theta$ obeys

$$ {}^{g}M_\theta(\mathcal{D}) = {}^{g}\overline{\theta}(\mathcal{D}) + \overline{Z}, \tag{10}$$

for some estimate ${}^{g}\overline{\theta}$ of ${}^{g}\theta$, such that the user-level sensitivity of ${}^{g}\overline{\theta}$ is $\Delta_{{}^{g}\overline{\theta}}$. Recall that the assumption that ${}^{g}M_\theta$ is a noise-adding mechanism is without loss of generality. Also, in (10), we have that $\overline{Z}$ is a length-d vector with $\overline{Z}_i \sim \text{Lap}\left(\Delta_{{}^{g}\overline{\theta}_i}/\epsilon\right)$, for each coordinate $i \in [d]$. Note that we work with the class of mechanisms that add Laplace noise tailored to the sensitivities of each grid, individually, since explicit computation of the user-level sensitivity of the vector f in (1) (across all grids) is quite hard, thereby implying the necessity of loose bounds on the amount of noise added, when this notion of user-level sensitivity is used.

Now, consider the mechanism M_θ that consists of the composition of the mechanisms ${}^{g}M_\theta$, over $g \in \mathcal{G}$, i.e., $M_\theta = ({}^{g}M_\theta : \ g \in \mathcal{G})$. In many settings of interest, a natural error metric for such a composition of mechanisms acting on different grids is the *largest worst-case* estimation error among all the grids.

Now, given a mechanism ${}^{g}M_\theta$ as in (10), we define its *worst-case* estimation error as

$$ {}^{g}E := \sum_{i\in[d]} \max_{\mathcal{D}\in\mathsf{D}} \left|{}^{g}\theta_i(\mathcal{D}) - {}^{g}\overline{\theta}_i(\mathcal{D})\right| + \mathbb{E}[\|\overline{Z}\|]. \tag{11}$$

Finally, we define the error metric E of the mechanism M_θ to be the *largest* worst-case estimation error among all the grids, i.e.,

$$ E := \max_{g\in\mathcal{G}} {}^{g}E. $$

We now describe our algorithm for reducing the privacy loss under composition, which makes use of a specialization of the definitions in this section to the case when the mechanisms ${}^{g}M_\theta$ are one of $M_\text{b} = {}^{g}M_\text{b}$ (corresponding to Baseline) or $M_\text{clip} = {}^{g}M_\text{clip}$ (corresponding to Clip).

5.2 An Algorithm for Clipping User Contributions

The algorithm discussed in this section results in a simple improvement of Theorem 2.2 that takes into account the structure of the queries. We mention that query-dependent composition results are also known for, say, histogram queries (see [17, Prop. 2.8]). Consider the Baseline mechanisms ${}^{g}M_\text{b} = \left[{}^{g}M_{\mu,\text{b}}, {}^{g}M_{\text{Var},\text{b}}\right]^\top$, as defined in Section 3.1, for estimating the statistics ${}^{g}\mu$ and ${}^{g}\text{Var}$, for each grid g of a given dataset, with $\mathsf{M}_\text{b} = ({}^{g}M_\text{b} : \ g \in \mathcal{G})$. Observe that initially, for any grid g, we have

$$\begin{aligned} {}^{g}E &= \mathbb{E}\left[|\text{Lap}(2\Delta_{{}^{g}\mu}/\varepsilon)|\right] + \mathbb{E}\left[|\text{Lap}(2\Delta_{{}^{g}\text{Var}}/\varepsilon)|\right] \\ &= \frac{2U\,{}^{g}m^\star}{\varepsilon\cdot\sum_{\ell\in{}^{g}\mathcal{L}}{}^{g}m_\ell} + \frac{2U^2\,{}^{g}m^\star\left(\sum_{\ell\in{}^{g}\mathcal{L}}{}^{g}m_\ell - {}^{g}m^\star\right)}{\varepsilon\cdot\left(\sum_{\ell\in{}^{g}\mathcal{L}}{}^{g}m_\ell\right)^2}, \end{aligned} \tag{12}$$

where the last equality follows from (5) and Proposition 3.1. As defined earlier, we have $E := \max_{g\in\mathcal{G}} {}^{g}E$. From Corollary 2.1, we notice that in order to improve the privacy loss upon composition, we must seek to reduce G_1, or the largest number of grids that any user "occupies". Our aim is to accomplish this reduction in such a manner as to not increase the worst-case error E[3].

5.2.1 The Iterative Procedure. Our algorithm proceeds in stages, at each stage suppressing *all* the contributions of those users that occupy the largest number of grids, in selected grids that these users occupy. Clearly, since the objective is to not increase E, for each such user, we suppress his/her contributions in that grid which has the smallest overall (that is the sum of errors due to bias and due to the noise added for privacy; see (11)) error *post suppression.* We emphasize that our algorithm, being iterative in nature, is not necessarily optimal in that it does not necessarily return the lowest possible privacy loss degradation factor for a fixed worst-case error E. Note also that while the worst-case error (over all grids) E is fixed at the start of the algorithm and is maintained as an invariant throughout its execution, the individual errors corresponding to each grid could potentially increase due to the suppression of user contributions. We let ${}^{g}E^{(0)} := {}^{g}E$, for each grid $g \in \mathcal{G}$.

For each step $t \geq 1$ in our algorithm, we pick the user(s) that occupy the largest number of grids. Define

$$ \mathsf{L}^{(t)} := \{\ell \in \mathcal{L} : \ G_\ell \geq G_j, \ \forall j \in \mathcal{L}\} $$

as the set of users in the first step of our algorithm that occupy the largest number of grids. The superscript '(t)' denotes the fact that the algorithm is in stage t of its execution. Recall from our assumption that $G_1 \geq G_j$, for any user $j \in \mathcal{L}$, and hence, in stage 1, we have user $1 \in \mathsf{L}^{(1)}$. We sort the users in $\mathsf{L}^{(t)}$ in increasing order of their indices, as $\ell_1 < \ell_2 < \ldots < \ell_{|\mathsf{L}^{(t)}|}$.

Now, for each user $\ell \in \mathsf{L}^{(t)}$, starting from user ℓ_1, we calculate the worst-case error that could result in each grid he/she occupies by potentially suppressing his/her contributions entirely. More precisely, for each grid $g \in \mathcal{G}_\ell$, we set ${}^{g}m_\ell = 0$, and recompute the values of $\Delta_{{}^{g}\mu}$ and $\Delta_{{}^{g}\text{Var}}$. In particular, following the definitions in Section 3.2, we note that after clipping in grid g, we have ${}^{g}\Gamma_\ell = 0$ and ${}^{g}\Gamma_{\ell'} = m_\ell$, for $\ell' \neq \ell$, with ${}^{g}\Gamma^\star = \max_{\ell'\in{}^{g}\mathcal{L}:\ \ell'\neq\ell} {}^{g}m_{\ell'}$. Thus, (5) and Proposition 3.1, can be used to compute the sensitivities of the new sample mean and sample variance in grid g, which we denote as ${}^{g}\Delta_{\mu_\text{clip}}(\ell)$ and ${}^{g}\Delta_{\text{Var}_\text{clip}}(\ell)$, respectively.

Moreover, such a clipping of the contributions of user $\ell \in \mathsf{L}^{(1)}$ in grid g introduces some *worst-case* clipping errors in the computation of μ and Var, which we call ${}^{g}E_\mu(\ell)$ and ${}^{g}E_\text{Var}(\ell)$, respectively. The exact magnitude of these clipping errors incurred can be computed using Theorems 4.1 and 4.2, using the same values of ${}^{g}\Gamma_{\ell'}$ and ${}^{g}\Gamma^\star$ as described above, for $\ell' \in {}^{g}\mathcal{L}$. Finally, following (11), we compute the overall worst-case error in grid g, post the suppression of the contributions of user ℓ as

$$ {}^{g}E(\ell) = {}^{g}E_\mu(\ell) + {}^{g}E_\text{Var}(\ell) + \frac{2\,{}^{g}\Delta_{\mu_\text{clip}}(\ell)}{\varepsilon} + \frac{2\,{}^{g}\Delta_{\text{Var}_\text{clip}}(\ell)}{\varepsilon}. \tag{13}$$

After computing the worst-case errors ${}^{g}E(\ell)$ that could result in each grid $g \in \mathcal{G}_\ell$ due to the potential suppression of the contributions of user ℓ in grid g, we identify one grid

$$ \mathrm{g}(\ell) \in \arg\min_{g\in\mathcal{G}_\ell} {}^{g}E(\ell) \tag{14}$$

[3] We mention that our algorithm can be executed with *any* bound E on the worst-case error of each grid and not just $E = \max_{g\in\mathcal{G}} {}^{g}E$.

and the corresponding error value ${}^{\mathrm{g}(\ell)}E(\ell)$. In the event that ${}^{\mathrm{g}(\ell)}E(\ell) \leq E$, where E is the original worst-case error, we proceed with clipping (or suppressing) all the contributions of user ℓ in grid $\mathrm{g}(\ell)$. In particular, we update ${}^{\mathrm{g}(\ell)}\mathcal{L} \longleftarrow {}^{\mathrm{g}(\ell)}\mathcal{L} \setminus \{\ell\}$ and $\mathcal{G}_\ell \longleftarrow \mathcal{G}_\ell \setminus \{\mathrm{g}(\ell)\}$. We recompute $G_\ell := |\mathcal{G}_\ell|$ and the above procedure, starting from (14), is then repeated for all users $\ell \in \mathsf{L}^{(t)}$.

Else, if ${}^{\mathrm{g}(\ell)}E(\ell) > E$, we reset ${}^{\mathrm{g}(\ell)}\Gamma_\ell$ to its original value at the start of the iteration and we halt the execution of the algorithm. We then return the value $K := \max_{\ell \in \mathcal{L}} G_\ell$ as the final privacy loss degradation factor. Note that, by design, the algorithm Clip-User maintains the worst-case error across grids as E, at every stage of its execution.

Algorithm 1 Clipping user contributions

```
procedure Clip-User($\mathcal{D}$)
    For each $g \in \mathcal{G}$, compute ${}^gE$ as in (12).
    Compute $E = \max_{g \in \mathcal{G}} {}^gE$.
    Set Halt ⟵ No and $t \longleftarrow 1$.
    while Halt = No do
        Compute $\mathsf{L}^{(t)} = \{\ell \in \mathcal{L} : G_\ell \geq G_j, \forall j \in \mathcal{L}\}$.
        for $\ell \in \mathsf{L}^{(t)}$ do
            for $g \in \mathcal{G}_\ell$ do
                Set ${}^g\Gamma_\ell = 0$.
                Compute error ${}^gE(\ell)$ as in (13).
            Pick $\mathrm{g}(\ell) \in \arg\min_{g \in \mathcal{G}_\ell} {}^gE(\ell)$.
            if ${}^{\mathrm{g}(\ell)}E(\ell) > E$ then
                Set Halt = Yes
                Reset ${}^g\Gamma_\ell$ to ${}^gm_\ell$, for all $g \in \mathcal{G}_\ell$.
                break
            else
                Restore ${}^g\Gamma_\ell$ to ${}^gm_\ell$, for all $g \in \mathcal{G}_\ell \setminus \{\mathrm{g}(\ell)\}$.
                Update $\mathcal{G}_\ell \longleftarrow \mathcal{G}_\ell \setminus \{\mathrm{g}(\ell)\}$ and ${}^g\mathcal{L} \longleftarrow {}^g\mathcal{L} \setminus \{\ell\}$.
                Compute $G_\ell$, for all $\ell \in \mathcal{L}$.
        if Halt = Yes then break
        else
            Set $t \longleftarrow t + 1$.
    Return $K \longleftarrow \max_{\ell \in \mathcal{L}} G_\ell$.
```

5.2.2 Post-Suppression Mechanism. Given the distribution $\{{}^gm_\ell\}$ of user contributions post the execution of Clip-User, we release user-level differentially private estimates of the sample means ${}^g\mu(\mathcal{D})$ and sample variances ${}^g\mathrm{Var}(\mathcal{D})$, for $g \in \mathcal{G}$, by using a version of the Clip mechanism for each grid, as discussed in Section 3.2. More precisely, for each grid g, we compute the values $\{{}^g\Gamma_\ell\}$ of user contributions post suppression, and release $M_{\text{clip, post}}(\mathcal{D}) = M_{\text{clip}}(\mathcal{D})$ as in (8). The following proposition then holds, similar to Corollary 2.1.

Proposition 5.1. *When ${}^g\epsilon = \epsilon$, for all $g \in \mathcal{G}$, we have that $M_{clip, post}$ is $K\epsilon$-DP, with a maximum worst-case error E over all grids.*

6 NUMERICAL RESULTS

In this section, we test the performance of Clip-User on synthetically generated datasets, via the privacy loss degradation $K = K_\epsilon$ obtained at the end of its execution. We first describe our experimental setup and then numerically demonstrate the improvements obtained in the privacy loss degradation factor by running Clip-User on these synthetic datasets.

6.1 Experimental Setup

Since this work concentrates on *worst-case* errors in estimation, it suffices to specify a dataset $\mathcal{D}$ by simply the collection $\{{}^gm_\ell : \ell \in \mathcal{L}, g \in \mathcal{G}\}$ of user contributions across grids. To this end, we work with the following distribution on the values $\{{}^gm_\ell\}$, which we believe is a reasonable, although much-simplified, model of real-world traffic datasets. We fix a number of grids $G = 12$ and a number of users $L = 2^{12} - 1 = 4095$.

(1) **User Occupancies**: We index the users $\ell \in \mathcal{L}$ from 1 to L. Any user $\ell \in [2^j : 2^{j+1} - 1]$ occupies (or, has non-zero contributions in) exactly $G - j = 12 - j$ grids, where $j \in [0 : G - 1]$. It is clear that in this setting, we have $G_1 \geq G_2 \geq \ldots \geq G_L$.
Now, consider any user ℓ that occupies k grids. We identify these k grids among the G overall grids by sampling a subset of $\mathcal{G}$ of cardinality k, uniformly at random.

(2) **Number of contributions**: For a user ℓ that occupies grids $g_1, \ldots, g_k$, for k fixed as above, we sample the number of his/her contributions in grid g_i, $i \in [k]$ as ${}^{g_i}m_\ell \sim \mathrm{Geo}(q)$, where $\mathrm{Geo}(q)$ denotes the geometric distribution with parameter $q \in [0, 1]$. In particular,
$$\Pr[{}^{g_i}m_\ell = m] = q \cdot (1 - q)^{m-1}, \; m \in \{1, 2, \ldots\}.$$

(3) **Scaling the maximum contributions**: For each grid $g \in \mathcal{G}$, we identify a single user $\ell \in \arg\max_{\ell' \in {}^g\mathcal{L}} {}^gm_{\ell'}$ and scale his/her number of contributions as ${}^gm_\ell \longleftarrow (1 + \gamma){}^gm_\ell$, for a fixed $\gamma > 0$.

We mention that Step 3 above is carried out to model most real-world datasets where there exists one user (or one vehicle, in our context) who contributes more samples than any other user, in each grid. Furthermore, note that the actual speed samples $\{{}^gS_\ell\}$ contributed by users across grids could be arbitrary, but these values do not matter in our analysis, since we work with the worst-case estimation errors.

6.1.1 Estimating Expected Privacy Loss Degradation. For a fixed γ, q, we draw 10 collections of (random) $\{{}^gm_\ell\}$ values. On each such collection of values, representing a dataset $\mathcal{D}$, we execute Clip-User and compute the privacy loss degradation factor K_ϵ for $\epsilon \in [0.1, 1]$. We mention that in our implementation of Clip-User, we refrain from clipping user contributions in that grid g = $\arg\min_{g \in \mathcal{G}} {}^gE$, for ${}^gE$ as in (12). As an estimate of the expected privacy loss degradation for the given γ, q parameters, we compute the Monte-Carlo average
$$\widehat{P_\epsilon} := \frac{1}{10} \sum_{i=1}^{10} K_\epsilon^{(i)} \epsilon,$$
where the index $i \in [10]$ denotes a sample collection of $\{{}^gm_\ell\}$ values as above, with $K_\epsilon^{(i)}$ denoting the privacy loss degradation returned by Clip-User for these values.

6.1.2 Improving Worst-Case Error. Now that we have (potentially) reduced the expected privacy loss degradation via the execution of Clip-User, while maintaining the worst-case error across grids as E, we discuss a simple strategy, drawing on [2], which seeks to reduce this worst-case error across grids. Let $\{{}^g\Gamma_\ell : g \in \mathcal{G}, \ell \in {}^g\mathcal{L}\}$ denote the distribution of user contributions across grids, for a fixed instantiation of user contributions as in Section 6.1, post suppression via Clip-User. Here, ${}^g\mathcal{L}$ denotes the set of users with non-zero contributions in grid g, *post* the execution of Clip-User.

In an attempt to reduce the worst-case error across grids further, we clip the contributions of *all* users in a grid g to some value $m \in [{}^g\Gamma_\star : {}^g\Gamma^\star]$, where ${}^g\Gamma_\star := \min_{\ell \in {}^g\mathcal{L}} {}^g\Gamma_\ell$ and ${}^g\Gamma^\star := \max_{\ell \in {}^g\mathcal{L}} {}^g\Gamma_\ell$. More precisely, for any fixed grid g, we pick the first ${}^g\overline{\Gamma}_\ell$ contributions of each user $\ell \in {}^g\mathcal{L}$, where ${}^g\overline{\Gamma}_\ell := \min\{{}^g\Gamma_\ell, m\}$, for some $m \in [{}^g\Gamma_\star : {}^g\Gamma^\star]$. This corresponds to using a *pseudo-user* creation-based clipping strategy, as mentioned in Section 3.2.

We then compute the sensitivities ${}^g\overline{\Delta}_{\mu_{\mathrm{clip}}}$ and ${}^g\overline{\Delta}_{\mathrm{Var}_{\mathrm{clip}}}$ of the resultant clipped estimators of the sample mean and variance, respectively, using (9) and Lemma 3.1 and the above values of $\{{}^g\overline{\Gamma}_\ell : \ell \in {}^g\mathcal{L}\}$. We also compute the clipping errors (or bias) introduced, which we call ${}^g\overline{E}_\mu$ and ${}^g\overline{E}_{\mathrm{Var}}$, using Theorems 4.1 and 4.2, with $\{{}^g\overline{\Gamma}_\ell : \ell \in {}^g\mathcal{L}\}$ corresponding to the clipped user contributions and $\{{}^g m_\ell\}$ corresponding to the original user contributions. Here, note that we use ${}^g\overline{\Gamma}_\ell = 0$ for those users $\ell \in \mathcal{L}$ with ${}^g m_\ell > 0$ and ${}^g\Gamma_\ell = 0$. We then set

$${}^g\overline{E}(m) := {}^g\overline{E}_\mu + {}^g\overline{E}_{\mathrm{Var}} + \frac{2\,{}^g\overline{\Delta}_{\mu_{\mathrm{clip}}}}{\epsilon} + \frac{2\,{}^g\overline{\Delta}_{\mathrm{Var}_{\mathrm{clip}}}}{\epsilon}$$

as the overall error post pseudo-user creation-based clipping in grid g, corresponding to a fixed value of m. Note that the errors involving the sensitivity terms correspond to a mechanism that adds Laplace noise to each of the clipped mean and variance functions, tuned to the sensitivities ${}^g\overline{\Delta}_{\mu_{\mathrm{clip}}}$ and ${}^g\overline{\Delta}_{\mathrm{Var}_{\mathrm{clip}}}$, respectively, with privacy loss parameter set to be $\epsilon/2$. We then compute

$${}^g\overline{E} := \min_{m \in [{}^g\Gamma_\star : {}^g\Gamma^\star]} {}^g\overline{E}(m),$$

and repeat these computations for each grid $g \in \mathcal{G}$. Finally, we set

$$\overline{E} = \overline{E}_\epsilon := \max_{g \in \mathcal{G}} {}^g\overline{E}$$

to be the new worst-case error across all grids.

As before, for our simulations, for a fixed γ, q, we draw 10 collections of (random) $\{{}^g m_\ell\}$ values. On each such collection of values, we execute Clip-User and the pseudo-user creation-based clipping strategy above for $\epsilon \in [0.1, 1]$. As an estimate of the expected worst-case error across grids *post* the execution of Clip-User, for the given γ, q parameters, we compute the Monte-Carlo average

$$\widehat{\overline{E}}_\epsilon := \frac{1}{10}\sum_{i=1}^{10} \overline{E}_\epsilon^{(i)},$$

where the index $i \in [10]$ denotes a sample collection of $\{{}^g m_\ell\}$ values as above, with $\overline{E}_\epsilon^{(i)}$ denoting the worst-case error across grids for these values.

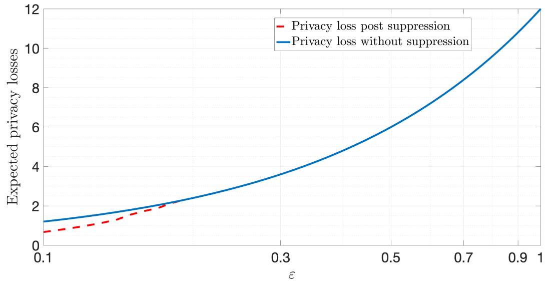

Figure 1: Plot of estimate $\widehat{P}_\epsilon$ of expected privacy loss, after execution of Clip-User, against the original privacy loss degradation $G\epsilon$. Here, $\gamma = 3$ and $q = 0.01$.

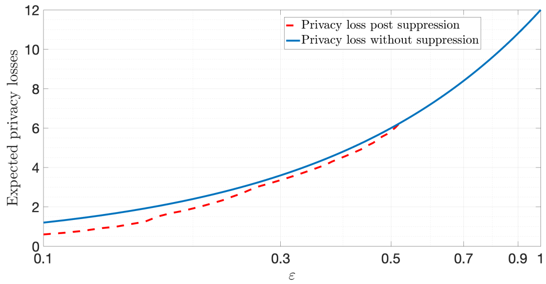

Figure 2: Plot of estimate $\widehat{P}_\epsilon$ of expected privacy loss, after execution of Clip-User, against the original privacy loss degradation $G\epsilon$. Here, $\gamma = 6$ and $q = 0.01$.

6.2 Simulations

Given the experimental setup described in the previous section, we now provide simulations that demonstrate the performance of Clip-User and the pseudo-user creation-based clipping strategy with regard to the expected privacy loss degradation and an expected worst-case error across grids.

Figures 1–3 show plots of the variation of the estimate $\widehat{P}_\epsilon$ of the expected privacy loss against the original privacy loss $G\epsilon = G_1\epsilon$ prior to the execution of Clip-User. The ϵ-axis is shown on a log-scale, here. From the plots, it is clear that for a fixed $q \in [0, 1]$, increasing γ improves the privacy loss degradation. Intuitively, a large value of γ leads to a large sensitivity of the unclipped mean and variance (and therefore a large worst-case error E); therefore, it is reasonable to expect many stages of Clip-User to execute before the algorithm halts, in this case.

Figures 4–6 show plots of the variation of the estimate of the expected worst-case error across grids $\widehat{\overline{E}}_\epsilon$ against the original worst-case error $E = E_\epsilon$ prior to the execution of Clip-User. Both the ϵ- and the error-axes are shown on a log-scale. Again, it is clear that for a fixed $q \in [0, 1]$, increasing γ leads to a larger difference between the original and the new worst-case errors, following similar intuition as that earlier.

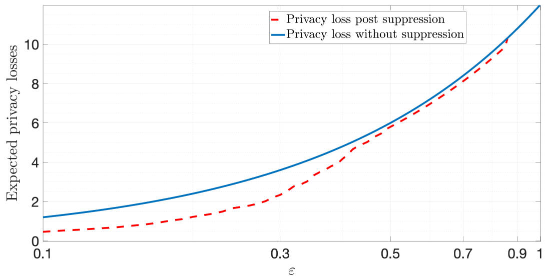

Figure 3: Plot of estimate $\widehat{P}_\epsilon$ of expected privacy loss, after execution of Clip-User, against the original privacy loss degradation $G\epsilon$. Here, $\gamma = 9$ and $q = 0.01$.

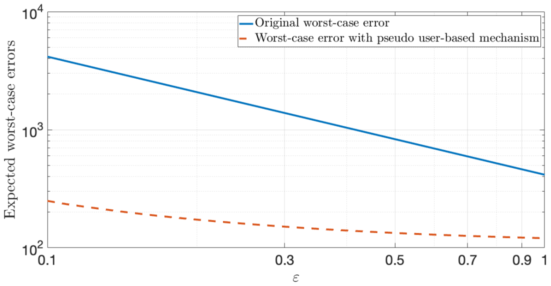

Figure 4: Plot of estimate $\widehat{\overline{E}}_\epsilon$ of the worst-case error across grids, after execution of Clip-User and the implementation of the pseudo-user creation-based clipping strategy, against the original worst-case error $E = E_\epsilon$. Here, $\gamma = 3$ and $q = 0.01$.

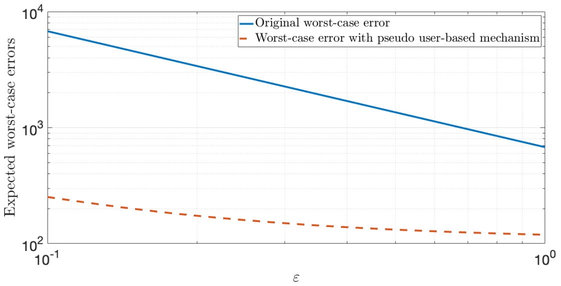

Figure 5: Plot of estimate $\widehat{\overline{E}}_\epsilon$ of the worst-case error across grids, after execution of Clip-User and the implementation of the pseudo-user creation-based clipping strategy, against the original worst-case error $E = E_\epsilon$. Here, $\gamma = 6$ and $q = 0.01$.

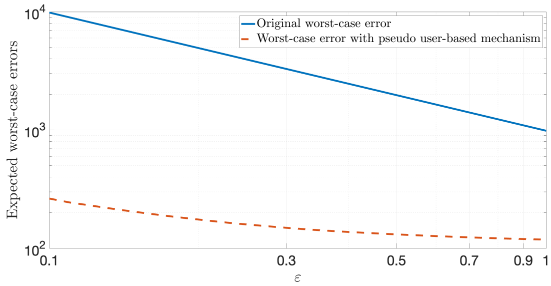

Figure 6: Plot of estimate $\widehat{\overline{E}}_\epsilon$ of the worst-case error across grids, after execution of Clip-User and the implementation of the pseudo-user creation-based clipping strategy, against the original worst-case error $E = E_\epsilon$. Here, $\gamma = 9$ and $q = 0.01$.

7 CONCLUSION

In this paper, we proposed an algorithm for improving the privacy loss degradation under the composition of user-level (pure) differentially private mechanisms that act on disjoint subsets of a dataset, in such a manner as to maintain the *worst-case* error in estimation over all such subsets. The basic idea behind our algorithm was the clipping of user contributions in selected subsets to improve the privacy loss degradation, while not increasing the worst-case estimation error. In particular, motivated by applications in the release of statistics of traffic data, we considered the design of such an algorithm for the release of the sample mean and variance of speed records in different grids in a city. A key component of the design of our algorithm was the explicit computation of the sensitivity of the sample variance function and the worst-case errors in estimation of the variance due to clipping selected contributions of users. We then presented numerical results evaluating the performance of our algorithm on synthetically generated datasets.

An interesting line of future research would be the extension of the techniques presented in this paper to the private release of other statistics of interest via their counts or histograms.

REFERENCES

[1] Uber Technologies Inc.. *H3: Hexagonal hierarchical geospatial indexing system.* https://h3geo.org/

[2] V. Arvind Rameshwar, Anshoo Tandon, Prajjwal Gupta, Novoneel Chakraborty, and Abhay Sharma. 2024. Mean Estimation with User-Level Privacy for Spatio-Temporal IoT Datasets. *arXiv e-prints*, Article arXiv:2401.15906 (Jan. 2024), arXiv:2401.15906 pages. https://doi.org/10.48550/arXiv.2401.15906 arXiv:2401.15906 [cs.CR]

[3] Rajendra Bhatia and Chandler Davis. 2000. A Better Bound on the Variance. *The American Mathematical Monthly* 107, 4 (2000), 353–357. https://doi.org/10.1080/00029890.2000.12005203 arXiv:https://doi.org/10.1080/00029890.2000.12005203

[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography* (2006), 265–284. https://doi.org/10.1007/11681878_14

[5] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407. https://doi.org/10.1561/0400000042

[6] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. 2010. Boosting and Differential Privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 51–60. https://doi.org/10.1109/FOCS.2010.12

[7] Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The Staircase Mechanism in Differential Privacy. *IEEE Journal of Selected Topics in Signal Processing* 9, 7 (2015), 1176–1184. https://doi.org/10.1109/JSTSP.2015.2425831

[8] Quan Geng and Pramod Viswanath. 2014. The optimal mechanism in differential privacy. In *2014 IEEE International Symposium on Information Theory*. 2371–2375. https://doi.org/10.1109/ISIT.2014.6875258

[9] Anand Jerry George, Lekshmi Ramesh, Aditya Vikram Singh, and Himanshu Tyagi. 2022. Continual Mean Estimation Under User-Level Privacy. *arXiv e-prints*, Article arXiv:2212.09980 (Dec. 2022), arXiv:2212.09980 pages. https://doi.org/10.48550/arXiv.2212.09980 arXiv:2212.09980 [cs.LG]

[10] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2017. The Composition Theorem for Differential Privacy. *IEEE Transactions on Information Theory* 63, 6 (2017), 4037–4049. https://doi.org/10.1109/TIT.2017.2685505

[11] Gautam Kamath and Jonathan Ullman. 2020. A primer on private statistics. *arXiv e-prints*, Article arXiv:2005.00010 (April 2020), arXiv:2005.00010 pages. https://doi.org/10.48550/arXiv.2005.00010 arXiv:2005.00010 [stat.ML]

[12] Daniel Asher Nathan Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. 2021. Learning with User-Level Privacy. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=G1jmxFOtY_

[13] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 111–125. https://doi.org/10.1109/SP.2008.33

[14] V. Pandurangan. 2014. *On Taxis and Rainbow Tables: Lessons for researchers and governments from NYC's improperly anonymized taxi logs.* https://blogs.lse.ac.uk/impactofsocialsciences/2014/07/16/nyc-improperly-anonymized-taxi-logs-pandurangan/

[15] Thomas Steinke. 2022. Composition of Differential Privacy & Privacy Amplification by Subsampling. *arXiv e-prints*, Article arXiv:2210.00597 (Oct. 2022), arXiv:2210.00597 pages. https://doi.org/10.48550/arXiv.2210.00597 arXiv:2210.00597 [cs.CR]

[16] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics* 25, 2–3 (1997), 98–110. https://doi.org/10.1111/j.1748-720x.1997.tb01885.x

[17] Salil Vadhan. 2017. *The Complexity of Differential Privacy.* Springer, Yehuda Lindell, ed., 347–450. https://link.springer.com/chapter/10.1007/978-3-319-57048-8_7

[18] C. Whong. 2014. *FOILing NYC's Taxi Trip Data.* https://chriswhong.com/open-data/foil_nyc_taxi/

A PROOF OF PROPOSITION 3.1

In this section, we shall prove Proposition 3.1.

Recall from the definition of user-level sensitivity in Section 2 that

$$\Delta_{\mathsf{Var}} = \max_{\mathcal{D}\sim\mathcal{D}'} \left|\mathsf{Var}(\mathcal{D}) - \mathsf{Var}(\mathcal{D}')\right|,$$

where Var is as in (4), and the notation $\mathcal{D} \sim \mathcal{D}'$ refers to the fact that $\mathcal{D}$ and $\mathcal{D}'$ are user-level neighbours, for $\mathcal{D}, \mathcal{D}' \in \mathsf{D}$. Moreover, without loss of generality, for the purpose of evaluating Δ_{Var}, we can assume that $\mathsf{Var}(\mathcal{D}') \leq \mathsf{Var}(\mathcal{D})$ in the expression for Δ_{Var}. Now, let

$$\mathsf{D}_{\max} = \left\{(\mathcal{D},\mathcal{D}') : (\mathcal{D},\mathcal{D}') \in \underset{\mathcal{D}\sim\mathcal{D}'}{\arg\max} \left|\mathsf{Var}(\mathcal{D}) - \mathsf{Var}(\mathcal{D}')\right|\right\}$$

be the collection of pairs of neighbouring datasets that attain the maximum in the definition of Δ_{Var}. In what follows, we shall exactly determine Δ_{Var} by identifying the structure of *one* pair $(\mathcal{D}_1, \mathcal{D}_2) \in \mathsf{D}_{\max}$ of neighbouring datasets.

Suppose that $\mathcal{D}_1, \mathcal{D}_2$ as above differ (exclusively) in the sample values contributed by user $k \in [L]$. Let $\left\{S_\ell^{(j)}\right\}$ denote the samples in dataset $\mathcal{D}_1$ and $\left\{\tilde{S}_\ell^{(j)}\right\}$ denote the samples in dataset $\mathcal{D}_2$. Let ν and $\tilde{\nu}$ be respectively the sample means of $\left\{S_\ell^{(j)}\right\}$ and $\left\{\tilde{S}_\ell^{(j)}\right\}$. Let $A := \left\{S_k^{(j)} : j \in [m_k]\right\}$ and $\tilde{A} := \left\{\tilde{S}_k^{(j)} : j \in [m_k]\right\}$ be the samples contributed by user k in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Further, let

$$\nu(A) := \frac{1}{m_k} \cdot \sum_{j=1}^{m_k} S_k^{(j)} \text{ and } \nu(\tilde{A}) := \frac{1}{m_k} \cdot \sum_{j=1}^{m_k} \tilde{S}_k^{(j)}$$

be the means of the samples in A and $\tilde{A}$, respectively. Similarly, let

$$\nu(A^c) := \frac{1}{\sum_{\ell\neq k} m_\ell} \cdot \sum_{\ell\neq k}\sum_{j=1}^{m_\ell} S_\ell^{(j)} \text{ and } \nu(\tilde{A}^c) := \frac{1}{\sum_{\ell\neq k} m_\ell} \cdot \sum_{\ell\neq k}\sum_{j=1}^{m_\ell} \tilde{S}_\ell^{(j)},$$

where we define A^c to be those samples contributed by the users other than user k in $\mathcal{D}_1$, and similarly, for $\tilde{A}^c$. By the definition of the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, we have that $A^c = \tilde{A}^c$ and hence $\tilde{\nu}(A^c) = \nu(A^c)$. Furthermore, the following lemma holds.

Lemma A.1. *There exists* $(\mathcal{D}_1, \mathcal{D}_2) \in \mathsf{D}_{max}$ *such that*

$$\nu(\tilde{A}) = \nu(A^c).$$

Furthermore, we can choose $\tilde{S}_k^{(1)} = \ldots = \tilde{S}_k^{(m_k)} = \nu(\tilde{A})$, *in* $\mathcal{D}_2$.

Proof. First, we write

$$\begin{aligned}\Delta_{\mathsf{Var}} &= \max_{\mathcal{D}\sim\mathcal{D}'} \left(\mathsf{Var}(\mathcal{D}) - \mathsf{Var}(\mathcal{D}')\right) \\ &= \max_{\mathcal{D}} \left(\mathsf{Var}(\mathcal{D}) - \min_{\mathcal{D}'\sim\mathcal{D}} \mathsf{Var}(\mathcal{D}')\right).\end{aligned}$$

Now, for a fixed dataset $\mathcal{D}$, consider $\mathsf{Var}(\mathcal{D}')$, for $\mathcal{D}' \sim \mathcal{D}$. Let $\tilde{X} \sim \text{Unif}(\tilde{A} \cup A^c)$ denote a uniformly distributed random variable that takes values in the set $\left\{\tilde{S}_\ell^{(j)}\right\}$. Then,

$$\begin{aligned}\mathsf{Var}(\mathcal{D}') &= \mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2\right] \\ &= \mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right] \Pr[\tilde{X} \in \tilde{A}] + \\ &\qquad \mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in A^c\right] \Pr[\tilde{X} \in A^c] \\ &\overset{(a)}{=} \mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right] \cdot \left(\frac{m_k}{\sum_\ell m_\ell}\right) + \\ &\qquad \mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in A^c\right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell}\right)\end{aligned}$$

Now, consider the term $\mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right]$ above. We can write

$$\begin{aligned}&\mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right] \\ &= \mathbb{E}\left[(\tilde{X} - \nu(\tilde{A}) + \nu(\tilde{A}) - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right] \\ &= \mathbb{E}\left[(\tilde{X} - \nu(\tilde{A}))^2 \mid \tilde{X} \in \tilde{A}\right] + (\tilde{\nu} - \nu(\tilde{A}))^2 + \\ &\qquad 2\,(\nu(\tilde{A}) - \tilde{\nu}) \cdot \mathbb{E}\left[(\tilde{X} - \nu(\tilde{A})) \mid \tilde{X} \in \tilde{A}\right].\end{aligned}$$

Clearly, since $\tilde{X} \sim \text{Unif}(\mathcal{D}')$, we have that conditioned on the event $\{\tilde{X} \in \tilde{A}\}$, we have that $\tilde{X}$ is uniform on $\tilde{A}$. Therefore, we obtain that $\mathbb{E}\left[(\tilde{X} - \nu(\tilde{A})) \mid \tilde{X} \in \tilde{A}\right] = 0$, implying that

$$\mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in \tilde{A}\right] = \mathbb{E}\left[(\tilde{X} - \nu(\tilde{A}))^2 \mid \tilde{X} \in \tilde{A}\right] + (\tilde{\nu} - \nu(\tilde{A}))^2. \tag{15}$$

By similar arguments, we obtain that

$$\mathbb{E}\left[(\tilde{X} - \tilde{\nu})^2 \mid \tilde{X} \in A^c\right] = \mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c\right] + (\tilde{\nu} - \nu(A^c))^2. \tag{16}$$

Substituting (15) and (16) into equality (a) above, we get that

$$\begin{aligned}&\mathsf{Var}(\mathcal{D}') \\ &= \left(\mathbb{E}\left[(\tilde{X} - \nu(\tilde{A}))^2 \mid \tilde{X} \in \tilde{A}\right] + (\tilde{\nu} - \nu(\tilde{A}))^2\right) \cdot \left(\frac{m_k}{\sum_\ell m_\ell}\right) + \\ &\qquad \left(\mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c\right] + (\tilde{\nu} - \nu(A^c))^2\right) \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell}\right).\end{aligned}$$

Now, observe that all the terms in equality (a) above are non-negative, and hence $\mathsf{Var}(\mathcal{D}')$ is minimized by setting $\tilde{\nu} = \nu(\tilde{A}) = \nu(A^c) = \tilde{S}_k^{(j)}$, for all $j \in [m_k]$. □

From the proof of the lemma above, we obtain that there exist datasets $(\mathcal{D}_1, \mathcal{D}_2) \in \mathsf{D}_{\max}$, such that

$$\mathsf{Var}(\mathcal{D}') = \mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c\right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell}\right).$$

Furthermore, for this choice of $\mathcal{D}_2$, we have $\tilde{S}_k^{(j)} = \nu(A^c)$, for all $j \in [m_k]$. The next lemma provides an alternative characterization of Δ_{Var}, using our choice of datasets $\mathcal{D}_1, \mathcal{D}_2$.

Lemma A.2. *We have that*

$$\Delta_{Var} = \max_{\mathcal{D}:\, S_\ell^{(j)}=\nu(A^c), \forall S_\ell^{(j)} \in A^c} Var(\mathcal{D})$$

Proof. Recall that

$$\Delta_{\mathsf{Var}} = \mathsf{Var}(\mathcal{D}_1) - \mathsf{Var}(\mathcal{D}_2),$$

with $\mathcal{D}_2$ chosen as in the discussion preceding this lemma. Thus, for random variables $\tilde{X} \sim \text{Unif}(\tilde{A} \cup A^c)$ and $X \sim \text{Unif}(A \cup A^c)$, we have

$$\begin{aligned}
&\Delta_{\mathsf{Var}} \\
&= \max_{\mathcal{D}} \left[\mathsf{Var}(\mathcal{D}) - \mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c \right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell} \right) \right] \\
&= \max_{\mathcal{D}} \left[\mathbb{E}\left[(X - \nu)^2 \right] - \mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c \right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell} \right) \right] \\
&\overset{(b)}{=} \max_{\mathcal{D}} \left[\mathbb{E}\left[(X - \nu)^2 \right] - \mathbb{E}\left[(X - \nu(A^c))^2 \mid X \in A^c \right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell} \right) \right],
\end{aligned}$$

where equality (b) follows from the fact that the distribution of $\tilde{X}$ conditioned on the event $\{\tilde{X} \in A^c\}$ is identical to that of X conditioned on the event $\{X \in A^c\}$. Hence,

$$\begin{aligned}
&\Delta_{\mathsf{Var}} \\
&= \max_{\mathcal{D}} \Bigg[\mathbb{E}\left[(X - \nu)^2 \mid X \in A \right] \Pr[X \in A] \\
&\qquad + \mathbb{E}\left[(X - \nu)^2 \mid X \in A^c \right] \Pr[X \in A^c] \\
&\qquad - \mathbb{E}\left[(\tilde{X} - \nu(A^c))^2 \mid \tilde{X} \in A^c \right] \cdot \left(1 - \frac{m_k}{\sum_\ell m_\ell} \right) \Bigg].
\end{aligned}$$

Now, observe that by arguments as in the proof of Lemma A.1,

$$\mathbb{E}\left[(X - \nu)^2 \mid X \in A^c \right] = \mathbb{E}\left[(X - \nu(A^c))^2 \mid X \in A^c \right] + (\nu - \nu(A^c))^2.$$

Now, since $\Pr[X \in A^c] = \left(1 - \frac{m_k}{\sum_\ell m_\ell} \right)$, we have by arguments made earlier, that

$$\begin{aligned}
\Delta_{\mathsf{Var}} &= \max_{\mathcal{D}} \Bigg[\mathbb{E}\left[(X - \nu)^2 \mid X \in A \right] \Pr[X \in A] + \\
&\qquad (\nu - \nu(A^c))^2 \Pr[X \in A^c] \Bigg] \qquad (17) \\
&= \max_{\mathcal{D}:\, S_\ell^{(j)}=\nu(A^c), \forall S_\ell^{(j)} \in A^c} \mathsf{Var}(\mathcal{D}),
\end{aligned}$$

thereby proving the lemma. □

Note that the maximization in the expression in Lemma A.2 is essentially over $\nu(A^c)$ and the variables $\left\{ S_j^{(\ell)} \right\} \in A$, with the constraint that $S_\ell^{(j)} = \nu(A^c)$, for all $S_\ell^{(j)} \in A^c$. It is easy to show that for a fixed choice of the cariables $\left\{ S_j^{(\ell)} \right\} \in A$, the expression in (17) is a quadratic function of $\nu(A^c)$, with a non-negative coefficient. Hence, the maximum over $\nu(A^c)$ of the expression in (17) is attained at a boundary point, i.e., at either $\nu(A^c) = 0$ or at $\nu(A^c) = U$. This observation immediately leads to a proof of Proposition 3.1.

Proof of Proposition 3.1. Recall from Lemma A.2 that

$$\Delta_{\mathsf{Var}} = \max_{\mathcal{D}:\, S_\ell^{(j)}=\nu(A^c), \forall S_\ell^{(j)} \in A^c} \mathsf{Var}(\mathcal{D}).$$

From the discussion preceding this lemma, consider the case when the maximum over $\nu(A^c)$ above is attained at $\nu(A^c) = 0$. The proof for the case when $\nu(A^c) = U$ follows along similar lines, and is hence omitted. In this case,

$$\Delta_{\mathsf{Var}} = \max_{S_\ell^{(j)} \in A:\, S_\ell^{(j)}=0, \forall S_\ell^{(j)} \in A^c} \mathsf{Var}(\mathcal{D}).$$

In this setting, $\nu \leq \frac{U m_k}{\sum_\ell m_\ell}$. Two possible situations arise: (i) when $\sum_\ell m_\ell > 2m^\star$, and (ii) when $\sum_\ell m_\ell \leq 2m^\star$. Consider the first situation. In this case, observe that $\nu \leq \frac{U m^\star}{\sum_\ell m_\ell} < U/2$. Further, from the Bhatia-Davis inequality [3], we have $\mathsf{Var}(\mathcal{D}) \leq \nu(U - \nu) =: b(\nu)$. Hence, for the range of ν values of interest, we have that $b(\nu)$ is strictly increasing in ν. Hence,

$$\begin{aligned}
\Delta_{\mathsf{Var}} &\leq \max_{\mathcal{D}:\, S_\ell^{(j)}=0, \forall S_\ell^{(j)} \in A^c} \nu(U - \nu) \\
&\leq \frac{U^2\, m^\star (\sum_\ell m_\ell - m^\star)}{(\sum_\ell m_\ell)^2},
\end{aligned}$$

with the inequalities above being achieved with equality when $S_k^{(j)} = U$, for all $j \in [m_k]$, and when $m_k = m^\star$. Next, consider the situation when $\sum_\ell m_\ell \geq 2m^\star$, and suppose that $\sum_\ell m_\ell$ is even. In this case, we have that $|A| \geq |A^c|$. For this setting, first note that

$$\Delta_{\mathsf{Var}} \leq \max_{\mathcal{D} \in \mathsf{D}} \mathsf{Var}(\mathcal{D}) = \max \mathsf{Var}(W) = \frac{U^2}{4},$$

for $W \sim \text{Unif}(A \cup A^c)$. To see why the above bound holds, note that for any bounded random variable $Y \in [0, U]$, we have that

$$\mathsf{Var}(Y) = \mathsf{Var}(Y - U/2) \leq U^2/4.$$

Furthermore, equality above is attained when all samples in A^c take the value 0 (which is in line with the case of interest where $\nu(A^c) = 0$) and $\frac{|A| - |A^c|}{2}$ samples in A take the value 0 and the remaining samples take the value U. This then results in exactly $\frac{\sum_\ell m_\ell}{2}$ samples being 0 and an equal number of samples being U, resulting $\Delta_{\mathsf{Var}} = U^2/4$.

Next, consider the case when $\sum_\ell m_\ell$ is odd. In this setting, it is not possible to ensure that equal number of samples (from $A \cup A^c$) are at 0 and U, thereby implying that the true value of $\mathsf{Var}(\mathcal{D})$, with $S_j^{(\ell)} = 0$, for all $S_j^{(\ell)} \in A^c$, in this case is smaller than $U^2/4$. We claim that in the case when the total number, $\sum_\ell m_\ell$, of samples is odd, the variance of a bounded random variable $Y \in [0, U]$ that takes values in $\left\{ S_j^{(\ell)} \right\}$ obeys

$$\mathsf{Var}(Y) \leq \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell m_\ell)^2} \right); \qquad (18)$$

furthermore, this bound is achieved when $\left\lceil \frac{\sum_\ell m_\ell}{2} \right\rceil$ samples take the value 0 and $\left\lfloor \frac{\sum_\ell m_\ell}{2} \right\rfloor$ samples take the value U. Modulo this claim, observe that in the case where $|A| \geq |A^c|$, the upper bound in (18) is achievable when $S_j^{(\ell)} = 0$, for all $S_j^{(\ell)} \in A^c$, by placing $\left\lceil \frac{|A| - |A^c|}{2} \right\rceil$ samples from A at the value 0 and the remaining samples at U.

We now prove the above claim. To this end, we first show that any sample distribution $\left\{S_j^{(\ell)}\right\}$ that maximizes the variance above must be such that $S_j^{(\ell)} \in \{0, U\}$, for all ℓ, j. For ease of reading, we let the samples $\left\{S_j^{(\ell)}\right\}$ be written as the collection $\{x_1, \ldots, x_n\}$, where $n = \sum_\ell m_\ell$. Now, we write

$$\max_{\mathcal{D}} \mathsf{Var}(Y) = \max_{x_1} \max_{x_2} \ldots \max_{x_n} \mathsf{Var}(Y). \tag{19}$$

Note that for fixed values of $x_1, x_2, \ldots, x_{n-1}$, the variance above is maximized when $x_n \in \{0, U\}$. To see why, let $\nu_{\sim n}$ denote the sample mean of the samples $x_1, \ldots, x_{n-1}$ and let $Y_{\sim n}$ denote the random variable that is uniformly distributed over the samples $x_1, x_2, \ldots, x_{n-1}$. By arguments as earlier, note that

$$\begin{aligned}
&\mathsf{Var}(Y)\\
&= \left(\frac{n-1}{n}\right) \cdot \left(\mathbb{E}\left[(Y - \nu_{\sim n})^2 \mid Y \neq x_n\right] + (\nu - \nu_{\sim n})^2\right) + \frac{1}{n}(\nu - x_n)^2\\
&= \left(\frac{n-1}{n}\right) \cdot \mathsf{Var}(Y_{\sim n}) + \frac{n-1}{n^2} \cdot (x_n - \nu_{\sim n})^2.
\end{aligned}$$

Clearly, the above expression is maximized, for fixed $x_1, x_2, \ldots, x_{n-1}$, by $x_n \in \{0, U\}$, depending on the value of $\nu_{\sim n}$. This argument can then be repeated iteratively over all $x_1, \ldots, x_n$, using (19).

Now, since all the samples in the collection $\{x_1, \ldots, x_n\}$ take a value of either 0 or U, all that remains is a maximization of $\mathsf{Var}(Y)$, given this constraint. Let k denote the number of samples taking the value 0 and let $n - k$ be the number of samples taking the value U. In this case, $\nu = \frac{(n-k)U}{n}$. Then,

$$\begin{aligned}
\frac{\mathsf{Var}(Y)}{U^2} &= \frac{k}{n} \cdot \left(\frac{n-k}{n}\right)^2 + \frac{n-k}{n} \cdot \left(\frac{k}{n}\right)^2\\
&= \frac{k(n-k)}{n^2}.
\end{aligned}$$

Clearly, when n is odd, the above expression is maximized when $\left\lceil \frac{n}{2} \right\rceil$ values are 0 and the remaining $\left\lfloor \frac{n}{2} \right\rfloor$ values are U, proving our earlier claim. □

B ON THE SENSITIVITIES UNDER A SPECIAL CLIPPING STRATEGY

In this section, we consider a special class of clipping strategies obtained by setting $\Gamma_\ell = \min\{m, m_\ell\}$, for some fixed $m \in [m_\star : m^\star]$. Clearly, here, we have $\Gamma^\star = m$ and $\Gamma_\star := \min_{\ell \in \mathcal{L}} \Gamma_\ell = m_\star$. Such a clipping strategy arises naturally in the design of user-level differentially private mechanisms based on the creation of pseudo-users [2, 9]. We show that for choices of m of interest, the sensitivities of the clipped sample mean and variance are at most those of their unclipped counterparts. In particular, for the sample mean, the following lemma was shown in [2]:

Lemma B.1 (Lemma III.1 in [2]). *For any $m \leq m^\star$, we have that $\Delta_\mu \geq \Delta_{\mu_{clip}}$.*

We now proceed to state and prove an analogous lemma that compares the sensitivities of $\mathsf{Var}_{\text{clip}}$ and Var. Before we proceed, observe that it is natural to restrict attention to those values of $m \in [m_\star : m^\star]$ that minimize the sensitivity of the clipped variance in 3.1. We first show that there exists a minimizer $m \in [m_\star : m^\star]$ that takes its value in the set $\{m_\ell\}_{\ell \in \mathcal{L}}$. Let $\Delta_{\mathsf{Var}_{\text{clip}}}(m) = \Delta_{\mathsf{Var}_{\text{clip}}}$, for a fixed m. To achieve this objective, we need the following helper lemma. For ease of exposition, we assume that $m_\star = m_1 \leq m_2 \leq \ldots \leq m_L = m^\star$. We also assume throughout that $L \geq 2$.

Lemma B.2. *$\Delta_{Var_{clip}}(m)$ is concave in m, for $m \in [m_t, m_{t+1}]$, for any $t \in [L-3]$, when $L \geq 3$.*

Proof. Fix an integer $t \in [L-1]$, for $L \geq 3$. Let $\alpha_1(m) := \frac{\Gamma_\ell^\star(\sum_\ell \Gamma_\ell - \Gamma_\ell^\star)}{(\sum_\ell \Gamma_\ell)^2}$, $\alpha_2(m) := \frac{1}{4}$, and $\alpha_3(m) := \frac{1}{4} \cdot \left(1 - \frac{1}{(\sum_\ell \Gamma_\ell)^2}\right)$.

Now, consider the setting where $t \leq L - 3$. In this case, observe that

$$\begin{aligned}
\sum_\ell \Gamma_\ell &= \sum_\ell \min\{m, m_\ell\}\\
&= \sum_{\ell=1}^{t} m_t + (L-t)m > 2m,
\end{aligned}$$

by our choice of t. This implies that for such values of t, we have $\Delta_{\mathsf{Var}_{\text{clip}}}(m) = U^2 \cdot \alpha_1(m)$, for all $m \in [m_t, m_{t+1}]$. Now, observe that we can write

$$\begin{aligned}
\alpha_1(m) &= \frac{m(c_1 + m(c_2 - 1))}{(c_1 + c_2 m)^2}\\
&= \frac{m}{c_1 + c_2 m} - \frac{m^2}{(c_1 + c_2 m)^2} =: a_1(m) + b_1(m),
\end{aligned}$$

for constants $c_1, c_2 > 0$ such that $c_1 + c_2 m = \sum_\ell \Gamma_\ell$. By direct computation, it is possible to show that

$$\frac{\mathrm{d}^2 a_1}{\mathrm{d}m^2} = \frac{-2c_1 c_2}{(c_1 + c_2 m)^3} < 0$$

and

$$\frac{\mathrm{d}^2 b_1}{\mathrm{d}m^2} = -2c_1 \cdot \left(\frac{c_1 + m(c_2 - 3)}{(c_1 + c_2 m)^4}\right) \leq 0,$$

since $c_1 + c_2 m = \sum_\ell \Gamma_\ell \geq 3m$, by our choice of t. Hence, for this case, we obtain that $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$ is concave in m. □

We are now ready to show that there exists a minimizer of the sensitivity $\Delta_{\mathsf{Var}_{\text{clip}}}$ that takes its value in $\{m_\ell\}$.

Lemma B.3. *There exists $m \in \arg\min_{m \in [m_\star, m^\star]} \Delta_{Var_{clip}}(m)$, such that $m \in \{m_\ell\}_{\ell \in \mathcal{L}}$.*

Proof. Suppose that $m \in [m_t, m_{t+1}]$, for some $t \in [L-1]$. We now argue that the value of $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$ cannot increase by setting m to $\arg\min_{m \in \{m_t, m_{t+1}\}} \Delta_{\mathsf{Var}_{\text{clip}}}(m)$. Indeed, note that if $t \in [L-3]$, by the concavity of $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$ from Lemma B.2, we obtain that a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$, for $m \in [m_t, m_{t+1}]$, occurs at a boundary point.

Now, consider the case when $t = L - 2$. In this case, observe that $\sum \Gamma_\ell - 2m = \sum_{\ell=1}^{L-2} m_\ell > 0$, if $L > 2$, and equals 0, if $L \leq 2$. Consider the first case when $L > 2$. In this setting, we have $\Delta_{\mathsf{Var}_{\text{clip}}}(m) = \alpha_1(m)$, for all $m \in [m_{L-2}, m_{L-1}]$. It is possible, by direct calculations, to show that when $m \in [m_{L-2}, m_{L-1}]$, we have

$$\frac{\mathrm{d}\alpha_1}{\mathrm{d}m} = \frac{-2mc_1}{(c_1 + c_2 m)^3},$$

for some constants $c_1, c_2 > 0$, thereby implying that α_1 is decreasing as a function of m, in this interval. Therefore, a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}$ occurs at a boundary point.

Next, consider the case when $L = 2$. In this setting, we have that $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$ equals either $\alpha_2(m)$ or $\alpha_3(m)$, for $m \in [m_{L-2}, m_{L-1}]$, when $\sum_\ell \Gamma_\ell$ is even or odd, respectively. Since $\alpha_2(m)$ is a constant and $\alpha_3(m)$ can be seen to be increasing in m in this interval, we obtain once again that a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}$ occurs at a boundary point.

Now, consider the case when $t = L - 1$. Observe that in this case, $\sum_\ell \Gamma_\ell - 2m = \sum_{\ell=1}^{L-1} m_\ell - m$ is decreasing as m increases from m_{L-1} to m_L. Hence, one of three possible cases can occur, each of which is dealt with in turn, below.

(1) $\sum_\ell \Gamma_\ell \le 2m$, for all $m \in [m_{L-1}, m_L]$: Clearly, in this case, we have that $\Delta_{\mathsf{Var}_{\text{clip}}}$ equals either α_2 or α_3, for $m \in [m_{L-2}, m_{L-1}]$, when $\sum_\ell \Gamma_\ell$ is even or odd, respectively. Since α_2 is a constant and $\alpha_3(m)$ is increasing with m in the interval of interest, we obtain that a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}$ occurs at a boundary point.

(2) $\sum_\ell \Gamma_\ell > 2m$, for all $m \in [m_{L-1}, m_L]$: Here, $\Delta_{\mathsf{Var}_{\text{clip}}} = \alpha_1$. Furthermore, we have that

$$\frac{\mathrm{d}\alpha_1}{\mathrm{d}m} = \frac{\sum_{\ell=1}^{L-1} m_\ell}{(m + \sum_{\ell=1}^{L-1} m_\ell)^2} > 0,$$

implying that $\Delta_{\mathsf{Var}_{\text{clip}}}$ is increasing in the interval of interest, hence showing that its minimizer occurs at a boundary point.

(3) $\sum_\ell \Gamma_\ell > 2m$, for $m \in [m_{L-1}, \overline{m}]$ and $\sum_\ell \Gamma_\ell \le 2m$, for $m \in (\overline{m}, m_L]$, for some $\overline{m} \in [m_{L-1}, m_L]$: Observe first that in this setting, we have that when $m = \overline{m}$,

$$\sum_\ell \Gamma_\ell = \overline{m} + \sum_{\ell=1}^{L-1} m_\ell = 2\overline{m},$$

by the definition of $\overline{m}$. In other words, we have $\overline{m} = \sum_{\ell=1}^{L-1} m_\ell$. Furthermore, for $m \in [m_{L-1}, \overline{m}]$, we have $\Delta_{\mathsf{Var}_{\text{clip}}}(m) = \alpha_1(m)$, while for $m \in (\overline{m}, m_L]$, we have $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$ equals $\alpha_2(m)$ or $\alpha_3(m)$, respectively, depending on whether $\sum_\ell \Gamma_\ell$ is even or odd. In the case when $\sum_\ell \Gamma_\ell$ is even, it can be verified that $\Delta_{\mathsf{Var}_{\text{clip}}}(\overline{m}) = \alpha_2(m) = 1/4$. Thus, using the fact that $\alpha_1(m)$ is increasing in m, we obtain that a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}$ occurs at a boundary point, when $\sum_\ell \Gamma_\ell$ is even.
Next, when $\sum_\ell \Gamma_\ell$ is odd, we have that $\alpha_3(m)$ is increasing in m for the interval of interest; we thus need only verify if for $L \ge 2$, we have

$$\alpha_3(\overline{m}) \ge \alpha_1(m_{L-1}).$$

Indeed, if the above inequality holds, we have that $\Delta_{\mathsf{Var}_{\text{clip}}}(\overline{m})$ is minimized at $m = m_{L-1}$, for $m \in [m_{L-1}, m_L]$. We can verify that the above inequality indeed holds, by a simple direct computation.

Hence, overall, we have that a minimizer of $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$, for $m \in [m_t, m_{t+1}]$ occurs at a boundary point, for all $t \in [L-1]$. □

Now that we have established that it suffices to focus on $m \in \{m_\ell\}_{\ell \in \mathcal{L}}$, we show that $\Delta_{\mathsf{Var}_{\text{clip}}}(m)$, for such values of m, is smaller than Δ_{Var}.

Lemma B.4. *When $\sum_\ell m_\ell$ is even, for $m \in \{m_\ell\}_{\ell \in \mathcal{L}}$, we have that $\Delta_{Var} \ge \Delta_{Var_{clip}}$.*

Proof. The proof proceeds by a case-by-case analysis. First, observe that if $m = m^\star$, we have that $\Gamma_\ell = m_\ell$, for all $\ell \in \mathcal{L}$, implying that $\Delta_{\mathsf{Var}} = \Delta_{\mathsf{Var}_{\text{clip}}}$. Hence, in what follows, we restrict attention to the case when $m \in [m_\star : m^\star - 1]$. Four possible scenarios arise:

(1) $\sum_\ell m_\ell \le 2m$: In this case, note that

$$\sum_\ell \Gamma_\ell < \sum_\ell m_\ell \le 2m < 2m^\star.$$

Hence, we have that $\Delta_{\mathsf{Var}} = U^2/4$, with $\Delta_{\mathsf{Var}_{\text{clip}}} = U^2/4$, if $\sum_\ell \Gamma_\ell$ is even, and $\Delta_{\mathsf{Var}_{\text{clip}}} = \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell \Gamma_\ell)^2}\right)$, if $\sum_\ell \Gamma_\ell$ is odd. Clearly, the statement of the lemma is true in this case.

(2) $\sum_\ell \Gamma_\ell \le 2m \le \sum_\ell m_\ell \le 2m^\star$: Here too, $\Delta_{\mathsf{Var}} = U^2/4$, with $\Delta_{\mathsf{Var}_{\text{clip}}} = U^2/4$, if $\sum_\ell \Gamma_\ell$ is even, and $\Delta_{\mathsf{Var}_{\text{clip}}} = \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell \Gamma_\ell)^2}\right)$, if $\sum_\ell \Gamma_\ell$ is odd. The lemma thus holds in this case as well.

(3) $2m < \sum_\ell \Gamma_\ell$: In this case, observe that

$$2m < \sum_\ell \Gamma_\ell < \sum_\ell m_\ell.$$

Hence, we have that $\Delta_{\mathsf{Var}_{\text{clip}}} = \frac{U^2 \ m(\sum_\ell \Gamma_\ell - m)}{(\sum_\ell \Gamma_\ell)^2}$. First, consider the case where $\sum_\ell m_\ell > 2m^\star$.
Again, without loss of generality, assume that $m_1 \ge m_2 \ge \ldots \ge m_L$. Let us define $\eta(t) := \frac{U^2 \ m^\star (t - m^\star)}{t^2}$, where $t \in (0, \infty)$. It is easy to show that $\frac{\mathrm{d}\eta}{\mathrm{d}t} < 0$, implying that $\eta(t)$ is decreasing in its argument t. Furthermore, if $m = m_1 = m^\star$, it is easy to see that $\Delta_{\mathsf{Var}_{\text{clip}}} = \Delta_{\mathsf{Var}}$. Now, suppose that $m = m_{j+1}$, for some $j \in [L-1]$, such that $m < m^\star = m_1$. Then, since $\sum_\ell m_\ell < (j+1)m_1 + \sum_{\ell=j+2}^{L} m_\ell =: c$, we have by the above analysis of the function η that

$$\Delta_{\mathsf{Var}} > \frac{U^2 m^\star((j+1)m_1 + c - m^\star)}{((j+1)m_1 + c)^2} =: \alpha$$

We next show that $\alpha \ge \Delta_{\mathsf{Var}_{\text{clip}}}(m)$. To this end, observe that

$$\Delta_{\mathsf{Var}_{\text{clip}}} = \frac{U^2 m((j+1)m + c - m)}{((j+1)m + c)^2}.$$

The result follows by explicitly computing $\alpha - \Delta_{\mathsf{Var}_{\text{clip}}}$ and arguing that this difference is non-negative, so long as $j \ge 2$, and hence, in particular, when $m < m^\star$. Hence, when $\frac{m_\star}{m^\star} > \frac{2}{L}$ and $m \in \{m_\ell\}$, we have that $\Delta_{\mathsf{Var}_{\text{clip}}} < \Delta_{\mathsf{Var}}$. For $\sum_\ell m_\ell \le 2m^\star$, we have that $\Delta_{\mathsf{Var}} = U^2/4 \ge \Delta_{\mathsf{Var}_{\text{clip}}}$, by a direct calculation.

(4) $\sum_\ell \Gamma_\ell \le 2m < 2m^\star \le \sum_\ell m_\ell$: We claim that such a situation cannot arise, for the given choice of Γ_ℓ, $\ell \in \mathcal{L}$. Indeed, observe that for $m \ne m^\star$, for $\sum_\ell \Gamma_\ell = \sum_\ell \min\{m, m_\ell\} \le 2m$ to hold, we must have that for some $\ell_0 \in \mathcal{L}$, the inequality $m_{\ell_0} > m$ holds, while $\sum_{\ell \ne \ell_0} m_\ell \le m$. This then implies that

$$\sum_\ell m_\ell = m_{\ell_0} + \sum_{\ell \ne \ell_0} m_\ell \le m$$
$$< m + m^\star < 2m^\star.$$

However, by assumption, we have that $2m^\star \le \sum_\ell m_\ell$, leading to a contradiction.

Putting together all the cases concludes the proof of the lemma. □

C PROOF OF THEOREM 4.2

In this section, we shall prove Theorem 4.2. Recall that we intend computing

$$E_{\mathsf{Var}} := \max_{\mathcal{D}\in \mathsf{D}} |\mathsf{Var}(\mathcal{D}) - \mathsf{Var}_{\mathsf{clip}}(\mathcal{D})|,$$

for fixed $\Gamma_\ell \in [0 : m_\ell]$, for $\ell \in \mathcal{L}$, with the assumption that $\sum_\ell \Gamma_\ell > 0$. For the case when $\Gamma_\ell = m_\ell$, for all $\ell \in \mathcal{L}$, it is clear that $\mathsf{Var}(\mathcal{D}) = \mathsf{Var}_{\mathsf{clip}}(\mathcal{D})$ and hence that $E_{\mathsf{Var}} = 0$. Hence, in what follows, we assume that there exists at least one user $\ell \in \mathcal{L}$ with $\Gamma_\ell < m_\ell$. Let $A := \left\{S_\ell^{(j)} : \ell \in \mathcal{L},\ j \in [\Gamma_\ell]\right\}$, and define

$$A^c := \left\{S_\ell^{(j)} : \ell \in \mathcal{L},\ j \in [\Gamma_\ell + 1 : m_\ell]\right\}.$$

Now, two cases can possibly arise: (i) when $|A| \le |A^c|$, and (ii) when $|A| > |A^c|$. Consider first case (i). Similar to the arguments made in the proof of Proposition 3.1, when $\sum_\ell m_\ell$ is even, we have that

$$E_{\mathsf{Var}} \le \max_{\mathcal{D}\in \mathsf{D}} \mathsf{Var}(\mathcal{D}) = \max \mathsf{Var}(X) = \frac{U^2}{4},$$

for $X \sim \mathrm{Unif}(A \cup A^c)$. Furthermore, equality above is attained when all samples in A take the value 0, and $\frac{|A^c|-|A|}{2}$ samples in A^c take the value 0 and the remaining samples take the value U. This then results in exactly $\frac{\sum_\ell m_\ell}{2}$ samples being 0 and an equal number of samples being U, resulting in a variance of $U^2/4$. Further, when $\sum_\ell m_\ell$ is odd, we have that

$$E_{\mathsf{Var}} \le \max_{\mathcal{D}\in \mathsf{D}} \mathsf{Var}(\mathcal{D}) = \max \mathsf{Var}(X) = \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell m_\ell)^2}\right),$$

with equality achieved when exactly $\left\lceil \frac{\sum_\ell m_\ell}{2} \right\rceil$ samples are 0 and $\left\lfloor \frac{\sum_\ell m_\ell}{2} \right\rfloor$ samples are U. This then gives rise to an exact characterization of E_{Var} when $\sum_\ell \Gamma_\ell \le \frac{\sum_\ell m_\ell}{2}$.

The setting of case (ii), when $|A| > |A^c|$, requires more work. However, the proof in this case is quite similar to the proof of Proposition 3.1. As in Appendix A, we define

$$\mu(A) := \frac{1}{\sum_\ell \Gamma_\ell} \cdot \sum_{\ell \in \mathcal{L}} \sum_{j=1}^{\Gamma_\ell} S_\ell^{(j)}$$

and

$$\mu(A^c) := \frac{1}{\sum_\ell (m_\ell - \Gamma_\ell)} \cdot \sum_{\ell \in \mathcal{L}} \sum_{j=\Gamma_\ell+1}^{m_\ell} S_\ell^{(j)}$$

as the sample means of the samples in the sets A and A^c, respectively. Further, let $\mu = \mu(\mathcal{D})$. The following lemma then holds.

Lemma C.1. *When $|A| > |A^c|$, we have that*

$$E_{Var} = \max_{\mathcal{D}\in \mathsf{D}} \left(Var(\mathcal{D}) - Var_{clip}(\mathcal{D})\right).$$

Proof.

$$\begin{aligned}
&\mathsf{Var}(\mathcal{D}) && (20)\\
&= \mathbb{E}_{X\sim \mathrm{Unif}(A\cup A^c)} \left[(X-\mu)^2\right]\\
&= \mathbb{E}\left[(X-\mu)^2 \mid X \in A\right] \cdot \mathbb{P}(A) + \mathbb{E}\left[(X-\mu)^2 \mid X \in A^c\right] \cdot P(A^c)\\
&= (\mu - \mu(A))^2 \cdot P(A) + \mathbb{E}\left[(X-\mu(A))^2 \mid X \in A\right] \cdot P(A) +\\
&\qquad (\mu - \mu(A^c))^2 \cdot P(A^c) + \mathbb{E}\left[(X-\mu(A^c))^2 \mid X \in A^c\right] \cdot P(A^c), && (21)
\end{aligned}$$

where we abbreviate $\Pr[X \in T]$ as $P(T)$, for some set $T \subseteq A \cup A^c$. The last equality holds for reasons similar to those in (15) and (16).

Next, note that

$$\begin{aligned}
&\mathsf{Var}_{\mathsf{clip}}(\mathcal{D}) && (22)\\
&= \mathbb{E}_{X'\sim \mathrm{Unif}(A)} \left[(X' - \mu(A))^2\right]\\
&= \mathbb{E}\left[(X-\mu(A))^2 \mid X \in A\right] P(A) + \mathbb{E}\left[(X-\mu(A))^2 \mid X \in A\right] P(A^c). && (23)
\end{aligned}$$

Putting together (21) and (23) and noting that, conditioned on the event $\{X \in A\}$, we have that X is uniform on the values in the set A, we get that

$$\begin{aligned}
&|\mathsf{Var}(\mathcal{D}) - \mathsf{Var}_{\mathsf{clip}}(\mathcal{D})| =\\
&\Big|(\mu - \mu(A))^2 P(A) + (\mu - \mu(A^c))^2 P(A^c) +\\
&\mathbb{E}\left[(X-\mu(A^c))^2 \mid X \in A^c\right] P(A^c) - E\left[(X-\mu(A))^2 \mid X \in A\right] P(A^c)\Big| && (24)
\end{aligned}$$

Now, consider a dataset $\overline{\mathcal{D}}$ such that the samples in A take the value 0 and the samples in A^c take the value U. Clearly, we have that

$$E_{\mathsf{Var}} \ge E_{\mathsf{Var}}(\overline{\mathcal{D}}) = \frac{U^2 \cdot |A| \cdot |A^c|}{(\sum_\ell m_\ell)^2} \qquad (25)$$

Furthermore, observe that

$$\begin{aligned}
&E_{\mathsf{Var}}\\
&= \max\left\{\max_{\mathcal{D}} \left(\mathsf{Var}(\mathcal{D}) - \mathsf{Var}_{\mathsf{clip}}(\mathcal{D})\right), \max_{\mathcal{D}} \left(\mathsf{Var}_{\mathsf{clip}}(\mathcal{D}) - \mathsf{Var}(\mathcal{D})\right)\right\}. && (26)
\end{aligned}$$

Now, from (24), note that

$$\begin{aligned}
\max_{\mathcal{D}} \left(\mathsf{Var}_{\mathsf{clip}}(\mathcal{D}) - \mathsf{Var}(\mathcal{D})\right) &\le \mathbb{E}\left[(X-\mu(A))^2 \mid X \in A\right] P(A^c)\\
&\le \frac{U^2 |A^c|}{4 \cdot \sum_\ell m_\ell} && (27)
\end{aligned}$$

By comparing (25) and (27), plugging back into (26), and noting that $|A| > |A^c|$, we obtain the statement of the lemma. □

Thus, from the above lemma and from (24), we obtain that when $|A| > |A^c|$,

$$\begin{aligned}
&E_{\mathsf{Var}} = (\mu - \mu(A))^2 P(A) + (\mu - \mu(A^c))^2 P(A^c) +\\
&\mathbb{E}\left[(X-\mu(A^c))^2 \mid X \in A^c\right] P(A^c) - E\left[(X-\mu(A))^2 \mid X \in A\right] P(A^c).
\end{aligned}$$

Now, clearly, we have that E_{Var} above is maximized by setting $X = \mu(A)$, when $X \in A$, or, in other words, setting $S_\ell^{(j)} = \mu(A)$, for all $\ell \in \mathcal{L}$ and $j \in [\Gamma_\ell]$. We thus obtain the following lemma:

Lemma C.2. *When $|A| > |A^c|$, we have that*

$$E_{Var} = \max_{\mathcal{D}:\, S_\ell^{(j)} = \mu(A), \forall S_\ell^{(j)} \in A} Var(\mathcal{D}).$$

Note the similarity between Lemma C.2 and Lemma A.2 in Appendix A. The proof of Theorem 4.2 is then immediate.

Proof of Theorem 4.2. Following on from Lemma C.2, by arguments analogous to those in the proof of Proposition 3.1 in Appendix A, we get that when $|A| > |A^c|$, $E_{\mathsf{Var}} = \frac{U^2 \cdot |A| \cdot |A^c|}{(\sum_\ell m_\ell)^2}$, which in turn equals $\frac{U^2 \cdot \sum_\ell \Gamma_\ell \cdot \sum_{\ell'} (m_{\ell'} - \Gamma_{\ell'})}{(\sum_\ell m_\ell)^2}$. The case when $|A| \leq |A^c|$ was already discussed earlier, wherein $E_{\mathsf{Var}} = \frac{U^2}{4}$, if $\sum_\ell m_\ell$ is even, and $E_{\mathsf{Var}} = \frac{U^2}{4} \cdot \left(1 - \frac{1}{(\sum_\ell m_\ell)^2}\right)$, if $\sum_\ell m_\ell$ is odd. □