Understanding Disclosure Risk in Differential Privacy with Applications to Noise Calibration and Auditing

Patricia Guerra-Balboa
Karlsruhe Institute of Technology, KASTEL SRL
patricia.balboa@kit.edu

Annika Sauer
Karlsruhe Institute of Technology, KASTEL SRL
annika.sauer@student.kit.edu

Héber H. Arcolezi
Inria, ÉTS Montréal
heber.hwang-arcolezi@etsmtl.ca

Thorsten Strufe
Karlsruhe Institute of Technology, KASTEL SRL
thorsten.strufe@kit.edu

ABSTRACT

Differential Privacy (DP) is widely adopted in data management systems to enable data sharing with formal disclosure guarantees. A central systems challenge is understanding how DP noise translates into effective protection against inference attacks, since this directly determines achievable utility. Most existing analyses focus only on membership inference—capturing only a threat—or rely on reconstruction robustness (ReRo). However, under realistic assumptions, we show that ReRo can yield misleading risk estimates and violate claimed bounds, limiting their usefulness for principled DP calibration and auditing.

This paper introduces reconstruction advantage, a unified risk metric that consistently captures risk across membership inference, attribute inference, and data reconstruction. We derive tight bounds that relate DP noise to adversarial advantage and characterize optimal adversarial strategies for arbitrary DP mechanisms and attacker knowledge. These results enable risk-driven noise calibration and provide a foundation for systematic DP auditing. We show that reconstruction advantage improves the accuracy and scope of DP auditing and enables more effective utility-privacy trade-offs in DP-enabled data management systems.

PVLDB Reference Format:
Patricia Guerra-Balboa, Annika Sauer, Héber H. Arcolezi, and Thorsten Strufe. Understanding Disclosure Risk in Differential Privacy with Applications to Noise Calibration and Auditing. PVLDB, 19(7): 1558 - 1571, 2026.
doi:10.14778/3801059.3801069

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/PatriciaBalboaKIT/Understanding-Risk-in-DP.

1 INTRODUCTION

Differential Privacy (DP) [24] and its distributed variant, local DP (LDP), have emerged as the de facto standard to mitigate privacy risk—that is, the extent to which a learning process allows sensitive information about participants to be inferred. DP aims to make participation as safe as not participating [23], and its privacy-utility trade-off is governed by the privacy budget ε (smaller values provide stronger guarantees) and by δ, which captures the probability mass of outcomes in which the guarantee may fail, weighted by the severity of their deviation from ε [53]. Despite this solid theoretical foundation, a central practical question remains: How do these formal parameters, especially ε, translate into concrete protection against real-world attacks? [54] This question is critical for calibrating ε: if set too high, sensitive information may be exposed; if too low, utility is unnecessarily compromised. Furthermore, understanding this relationship is essential for DP auditing, which aims to empirically estimate privacy [37], test the tightness of DP mechanisms [56], and detect bugs [65].

Motivated by its applications in noise calibration and auditing, there is growing interest in the data management community in risk assessment for DP mechanisms [7, 12, 16, 18]. Significant progress has been made in connecting DP to the risk of *membership inference attacks* (MIAs) [12, 26, 36, 70], even enabling direct noise calibration for desired MIA risk levels [44] without explicitly choosing ε. However, MIAs capture only one aspect of privacy risk and may be less relevant in deployments such as census data releases. In particular, *attribute inference attacks* (AIAs) [70], which can expose sensitive information even when membership is public [9], remain less understood. Recently, *data reconstruction attacks* (DRAs) [9] were proposed as a unifying framework subsuming both MIAs and AIAs, while also accounting for partial or imperfect reconstruction, e.g., revealing a car's license plate may suffice to compromise privacy even if the background is inaccurate.

Balle et al. [9] introduced the first metric for DRAs, *reconstruction robustness* (ReRo), providing a pioneering unified view of DP attack resilience. ReRo was foundational, but has limitations as a comprehensive adversarial metric. First, ReRo and existing bounds [9, 34] assume attackers have no target-specific auxiliary knowledge, ignoring partial information such as demographic attributes or social media data—information that real-world attacks often exploit [20, 55, 63]. We empirically confirm this limitation: when target-specific auxiliary information is available, the empirical ReRo exceeds the existing ReRo bounds (see Figure 3). Second, ReRo is a success probability, which penalizes mechanisms for providing global statistical knowledge—the end goal of data release—and incorrectly accounts for success from background knowledge or statistical imputation as participation risk [15, 43], leading to unnecessary utility loss when used for noise calibration (Figure 2).

We address such limitations by introducing *reconstruction advantage* (RAD), which extends advantage metrics to the unifying

Proceedings of the VLDB Endowment, Vol. 19, No. 7 ISSN 2150-8097.
doi:10.14778/3801059.3801069

DRA framework. RAD overcomes ReRo's limitations, naturally incorporating auxiliary knowledge and avoiding risk overestimation. We establish tight bounds linking DP noise to RAD, enabling noise injection calibrated to a participant's true risk of information disclosure. Specifically, we provide: (i) a worst-case bound independent of the attacker's auxiliary knowledge (Theorem 4.2), and (ii) an auxiliary-dependent, universally tight bound (Theorem 4.3). To assess tightness, we construct and prove the optimal attack strategy for any reconstruction goal, auxiliary knowledge, and mechanism—which also serves as a practical tool for DP auditing.

Theorem 4.3 is universally tight and cannot be further improved. However, it requires full knowledge of the mechanism $\mathcal{M}$, limiting its applicability in auditing external software. While Theorem 4.2 can serve as a fallback in such scenarios, it may strongly overestimate risk when no auxiliary information is available. To address this, we provide closed-form, black-box upper bounds for RAD without auxiliary knowledge (i.e., when the entire target record is considered secret, as in [6, 9, 34]) and for the case of perfect reconstruction, which is particularly relevant for categorical data where sensitive attributes (e.g., diseases, political opinions, or religious beliefs) cannot be partially reconstructed [28, 29]. All our bounds substantially reduce the required noise compared to existing ReRo bounds, and we validate these improvements experimentally.

These results provide the theoretical foundation for practical DP auditing. Modern DP systems deployed in industry [27], government [2], and data-processing pipelines [52] still lack general-purpose tools for quantifying real-world privacy leakage. Existing auditing tools either focus on a narrow attack class (often MIAs) [6, 37, 50, 56, 65] or rely on learning-based strategies requiring extensive tuning without mechanism-independent guarantees [48]. RAD fills this gap, offering a principled, mechanism-agnostic characterization of reconstruction risk. Building on our novel bounds, we introduce a RAD-based auditing framework that generalizes beyond prior tools [6, 21], capturing all reconstruction risks and providing more accurate, actionable privacy assessments. While our auditing framework is general in scope, in this paper we instantiate it for LDP and address key limitations of the state-of-the-art tool, LDP Auditor [6]. Unlike LDP Auditor, which relies on perfect reconstruction without target-specific auxiliary knowledge—and thus misses important threats such as AIAs—our method is both more general and produces *tighter empirical estimates* of the privacy budget for all the tested LDP mechanisms as demonstrated in our empirical study (see Figure 5).

Our contributions are summarized as follows:

- We empirically show that ReRo and its existing bounds fail to account for imputation-based success and target-specific auxiliary knowledge, limiting applicability.
- We introduce *Reconstruction Advantage (RAD)* as a consistent, unifying risk metric that naturally incorporates auxiliary knowledge.
- We establish tight worst-case and auxiliary-dependent bounds for RAD, along with black-box bounds for attackers lacking auxiliary knowledge.
- We construct the optimal attack strategy for any reconstruction goal, mechanism, and prior distribution, proving its optimality and demonstrating empirical utility for auditing.
- We propose a RAD-based DP auditing framework that provides broader threat analyses and more accurate privacy-budget estimates than existing LDP auditing techniques.

We provide detailed proofs, as well as additional experiments and computations in the long version of this paper (arXiv:2603.12142).

2 BACKGROUND

In this section, we introduce the relevant concepts for this work and present the notation used throughout the manuscript.

Differential Privacy. We assume each record $z \in \mathcal{Z}$ to be drawn independently from an underlying prior distribution $\mathcal{Z} \sim \pi$. Let $\mathcal{D}(\Theta)$ denote the space of probability distributions over the output space Θ. We consider a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ which, given an input database $D \in \mathcal{Z}^n$, produces a global output (e.g., an aggregate statistic or a trained model) $\theta \in \Theta$ with probability/density function $p_{\mathcal{M}}(\theta \mid D)$. In this context, DP is formalized as follows:

Definition 2.1 ([24]). A mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ is (ε, δ)-differentially private if for all $S \subseteq \Theta$ and for every pair of datasets $D_0, D_1 \in \mathcal{Z}^n$ such that $d_H(D_0, D_1) \leq 1$:

$$\Pr(\mathcal{M}(D_0) \in S) \leq \mathrm{e}^{\varepsilon} \Pr(\mathcal{M}(D_1) \in S) + \delta$$

where $d_H(D_0, D_1)$ denotes the Hamming distance [49].

If $\delta = 0$ we speak of *pure* DP (ε-DP). If $n = 1$, i.e., $\mathcal{M}$ takes as input a single data record $z \in \mathcal{Z}$, we obtain *Local Differential Privacy* (LDP). LDP is a rigorous and increasingly relevant privacy model in which data is randomized on the client side before being transmitted to a data collector [25]. Consequentially, it is especially suitable for privacy-sensitive applications such as telemetry and location-based services where no trusted data curator is considered [27].

The privacy budget ε determines how closely the probabilities of observing the same output on databases D_0 and D_1 must align, hence bounding their statistical "indistinguishability". A smaller ε provides stronger privacy guarantees but typically comes at the cost of utility [25]. The parameter δ allows certain violations of ε-DP while characterizing how likely such failures are to occur and the degree of such failures. Consequently, we aim to parameterize the attack performance based on the privacy parameters.

Many real-world deployments apply multiple DP mechanisms sequentially [14, 19]. By DP's adaptative composition property, the total privacy loss is determined by the parameters of the individual mechanisms [42]. Formally, given $[T] = \{1, \ldots, T\}$, for each $i \in [T]$, let $\overline{\Theta}_{i-1} = \prod_{j=1}^{i-1} \Theta_j$ denote the space of previous outputs, and define $\mathcal{M}_i\colon \mathcal{Z}^n \times \overline{\Theta}_{i-1} \to \Theta_i$. The *$T$-fold composed* mechanism is $\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D, Y_1), \ldots, \mathcal{M}_T(D, Y_{T-1}))$, where $Y_i = (\mathcal{M}_1(D), \ldots, \mathcal{M}_i(D, Y_{i-1}))$ denotes the first i outputs. Dwork et al. [24] established the first general bound on the privacy loss under T-fold adaptive composition: Composing (ε, δ)-DP mechanisms yields to $(T\varepsilon, T\delta)$-DP. Subsequent refinements led to tighter composition bounds as presented in [42].

Differential Privacy and Attack Resilience. Following previous work we consider for any target record z an *informed adversary* [9] with access to: the fixed dataset $D_- = D \backslash \{z\}$, the distribution of data records π, the output θ of the model trained on $D_z = D_- \cup \{z\}$, the mechanism $\mathcal{M}$, and optional target-specific auxiliary knowledge

$a(z)$ about target record z. We adopt this adversary model because, under the assumption that records are independently drawn from π, bounding the performance of such an attacker also bounds the performance of any attacker with less information [9].

Our analysis focuses on DRAs, where the adversary's goal is to correctly reconstruct completely or partially the target record z, potentially given auxiliary knowledge $a(z) \in aux$ about the target. DRAs cover AIAs and MIAs as particular cases [9]: In an MIA, the attacker knows the entire target record $a(z) = z$ and seeks only to infer its participation in the dataset. In an AIA, records are structured as $z = (x, y)$, $a(z) = x$ is considered public and the attacker aims to perfectly reconstruct the sensitive attribute y. More generally, in a DRA setting, it is natural to assume access to target-specific auxiliary knowledge. For example, when reconstructing a license plate number from a target's car image, the attacker may already know the color of the car. Hence, DRAs cover the broad range of commonly discussed privacy risks, including MIAs and AIAs as a particular instance [9]. Formally, a DRA, denoted by $A\colon \Theta \times aux \to \mathcal{Z}$ uses the output of a DP mechanism $\theta \sim \mathcal{M}(D)$ and the target auxiliary information to produce a candidate $\tilde{z} = A(\theta, a(z))$. Note that, in case of composing several mechanisms, we consider the final output after the whole process.

The attack is considered successful if the output is similar enough (according to a success threshold η) to the real input z: $\ell(\tilde{z}, z) \leq \eta$. The error function ℓ depends on the context, for instance, in a classic AIA, given $z = (x, y)$ we define $\phi(z) = y$ and $\ell(\tilde{z}, z) = 0$ if $\phi(\tilde{z}) = \phi(z)$ and one otherwise. In a MIA, ℓ is the characteristic function such that $\ell(\tilde{z}, z) = 0$ when $\tilde{z} = z$ and one otherwise. However, it may be sensitive enough to partially reconstruct the target, for instance, the image domain, even if not all pixels are correct. In this case, we may gather sensitive information such as the action performed in the image and therefore ℓ is chosen as an image-specific metric, such as the Learned Perceptual Image Patch Similarity (LPIPS) [9]. Given the error function ℓ and the threshold η, we define the *success set* of a target z as $S_\eta(z) = \{z' \in \mathcal{Z}\colon \ell(z, z') \leq \eta\}$.

After defining a DRA, the question of how to evaluate its performance arises. For the particular cases of AIA and MIAs, the current literature [32, 70] agrees on the following metric:

Definition 2.2 (Adapted from [70]). Given π the distribution of data records and $\mathcal{M}, \phi(z), a(z), A$ as defined above, the *attribute advantage*, Adv_{AIA}, is defined as

$$\Pr_{\substack{z_0\sim\pi\\ \theta\sim\mathcal{M}(D_{z_0})}}[A(\theta, a(z_0)) = \phi(z_0)] - \Pr_{\substack{z_0,z_1\sim\pi\\ \theta\sim\mathcal{M}(D_{z_1})}}[A(\theta, a(z_0)) = \phi(z_0)]\,.$$

The attribute advantage measures the adversary's gain in correctly inferring a sensitive attribute $\phi(z)$ when the record is in the input dataset $z_0 \in D$, compared to when it is drawn from the underlying distribution π. The second term in Definition 2.2 corrects for cases where the attribute could be inferred even without the record being in the database (e.g., through imputation [40]).

The current proposed performance metric for general DRAs [9] does not define an advantage but instead only accounts for the success probability of an attack that has as input solely the output of the DP mechanism and the known dataset D_-, ignoring any possible target-specific auxiliary knowledge:

Definition 2.3 (ReRo [9]). Let π be a prior over $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ a error function. Mechanism $\mathcal{M} : \mathcal{Z}^n \to \mathcal{D}(\Theta)$ is (η, γ)-reconstruction robust with respect to π, ℓ if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction adversary $A : \Theta \to \mathcal{Z}$,

$$\Pr_{Z\sim\pi, \theta\sim\mathcal{M}(D_Z)}[\ell(Z, A(\theta)) \leq \eta] \leq \gamma.$$

The first bound for ReRo under ε-DP was given by [9]:

$$\gamma \leq \kappa^+_{\pi,\ell}(\eta)\mathrm{e}^\varepsilon, \tag{1}$$

where $\kappa^+_{\pi,\ell}(\eta) = \sup_{z_0} \Pr_{Z\sim\pi}[\ell(z_0, Z) \leq \eta]$. Intuitively, $\kappa^+_{\pi,\ell}(\eta)$ represents the success probability of an oblivious attack that always selects the most likely reconstruction under the prior π.

Recent work [34] refined this bound using f-DP [22], a characterization of DP that captures the exact statistical indistinguishability between neighbors through the functional f. Formally,

Definition 2.4 ([43]). Let $f\colon [0, 1] \to [0, 1]$ be a continuous, convex, non-increasing function such that $f(x) \leq 1 - x$. A mechanism $\mathcal{M}$ satisfies f-DP if for all $D_0, D_1 \in \mathcal{Z}^n$ such that $d_H(D_0, D_1) \leq 1$ and all post-processing algorithms $A\colon \mathrm{Range}(\mathcal{M}) \to \mathcal{D}(\{0, 1\})$,

$$\Pr(A(\mathcal{M}(D_0)) = 1) \leq 1 - f\big(\Pr(A(\mathcal{M}(D_1)) = 1)\big).$$

Here, f is known as a *trade-off function* [22], named for its interpretation in the context of hypothesis testing. Specifically, consider A as a test of H_0: the input is D_0 vs. H_1: the input is D_1, applied to the output of $\mathcal{M}$. Then $\Pr(A(\mathcal{M}(D_0)) = 1)$ is the significance level and $\Pr(A(\mathcal{M}(D_1)) = 1)$ is the power of the test. Under this interpretation, for a given significance level, f bounds the maximum achievable power. When f is the trade-off function between two normal distributions with different means, namely $f(\alpha) = \Phi\big(\Phi^{-1}(1 - \alpha) - \mu\big)$, where Φ denotes the standard normal CDF, the resulting notion is known as *Gaussian DP* (μ-GDP).

The f-DP framework facilitates the computation of quantities such as the *total variation* distance:

Definition 2.5. A mechanism $\mathcal{M}$ has total variation at most $\mathrm{TV}(\mathcal{M})$ if, for all neighboring datasets D_0, D_1,

$$\sup_{S\subseteq\Theta} |\Pr(\mathcal{M}(D_0) \in S) - \Pr(\mathcal{M}(D_1) \in S)| \leq \mathrm{TV}(\mathcal{M}).$$

For any $\mathcal{M}$ satisfying (ε, δ)-DP, its TV is bounded [42] as

$$\mathrm{TV}(\mathcal{M}) \leq \max_{\alpha\in[0,1]} \big(1 - f(\alpha) - \alpha\big) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}. \tag{2}$$

Both f-DP and TV are preserved under composition. Specifically, the T-fold composition of an f-DP mechanism satisfies $f^{\otimes T}$-DP, where $f \otimes f$ denotes the trade-off function $T(P \times P, Q \times Q)$ for $f = T(P, Q)$. For instance, if a mechanism is μ-GDP, then its T-fold composition is $(\mu\sqrt{T})$-GDP [22]. Moreover, if $\mathrm{TV}(\mathcal{M}_i) = \Delta$, then the T-fold composition satisfies $\mathrm{TV}(\mathcal{M}) \leq 1 - (1 - \Delta)^T$ [30]. This bound can be sharpened to $\max_\alpha\big(1 - f^{\otimes T}(\alpha) - \alpha\big)$ when f is known.

Hayes et al. [34] present the first bound for any f-DP mechanism:

$$\gamma \leq 1 - f(\kappa^+_{\pi,\ell}(\eta)). \tag{3}$$

which they showed empirically nearly tight for DP-SGD, the most known DP algorithm for private learning [1].

3 REVIEW OF THE RELATED WORK

In this section, we review relevant prior work on measuring the effective attack resilience of DP mechanisms for calibration and auditing, discussing novel insights and gaps that motivate our work.

Attack-Based DP Noise Calibration. Several recent studies [12, 17, 44] demonstrate that calibrating DP noise based on resilience to specific attacks can significantly improve utility. Such approaches, however, primarily target MIAs, which leads to unnecessary utility degradation without offering meaningful privacy benefits when membership is public or considered non-sensitive [9].

Beyond MIAs, privacy concerns often involve AIA, where the adversary aims to infer sensitive attributes of individuals from released data [38, 59]. A common metric for evaluating such attacks is the attribute advantage [70]. Existing works that provide theoretical bounds for AIAs either analyze specific attack strategies [70] or adopt more general DRA frameworks [9, 32]. Within the latter, the notion of ReRo has emerged as the metric for measuring the risk of DRAs, under which attribute inference can be modeled as a special case [9]. Moreover, Equation (1) [9] and Equation (3) [34] provide ReRo-based DP noise calibration methods.

A note on limitations of ReRo. Balle et al.'s pioneering work [9] introduced ReRo and linked it to DP, providing a framework to assess the risks of DRAs and enabling risk analysis beyond MIAs. ReRo is suitable when the adversary's reconstruction capability is entirely based on the participation of the record, yet extending it to broader settings introduces significant limitations.

A general-purpose risk metric would be expected to cover all relevant attack scenarios. However, ReRo does not formally account for the impact of target-specific auxiliary knowledge, hence excluding MIAs, AIAs and targeted DRAs as introduced in Section 2. Formally, the attack considered in [9] (see Definition 2.3 for details), only has access to the mechanism output $\mathcal{M}(D)$, i.e., $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, implying that $\Pr(A(\mathcal{M}(D), a(z)) \in S) = \Pr(A(\mathcal{M}(D), a(z')) \in S)$ for any pair of possible targets z, z' and output set S. Under this assumption, the attacker A cannot adapt its strategy to a specific target z. This choice fundamentally prevents assessing the risk of MIA and AIA, as they use full or partial knowledge of some target records. This is a relevant limitation since most real-world privacy attacks historically exploit publicly available information about the target [20, 55, 63]. Moreover, we show in Sections 4 and 7 several attacks that leverage target-specific auxiliary knowledge, and their success highly depends on it.

All formal bounds connecting ReRo and DP were proven under this restrictive exclusion. The requirement that the attack depends only on $\mathcal{M}(D)$—ignoring target-specific information—is critical to establishing both Equations (1) and (3). This is not merely a theoretical limitation: we show in Section 7 that these bounds do not hold for attacks that exploit target-specific knowledge against well-known mechanisms such as DP-SGD.

A direct extension of ReRo to targeted attacks $A(\theta, a(z))$ fails: Not only do the original bounds no longer hold, but the metric also collapses to a substantial overestimation of risk due to imputation and background knowledge. For instance, the trivial MIA, $A(\theta, z) = z$, has success probability 1, which ReRo would interpret as a catastrophic privacy risk, even though no actual leakage occurs. This is not a negligible edge case; it has caused misleading overestimation of risk in black-box attacks on classification models [40], where much of the reported success arose from data imputation rather than exploiting the mechanism's output. Such overestimation obscures the true leakage and can lead to unnecessary utility loss when ReRo is used to calibrate noise in DP.

Even under the original assumption that the attacker has no target-specific knowledge, ReRo still overestimates risk, as we discussed in our preliminary work [32]. The mechanism output $\mathcal{M}(D)$ inherently reveals distributional information and population-level statistics, which are the primary goals of any learning process. This information can be used to perform imputation and infer attributes of individual records—even those not in D—with high accuracy, particularly when strong correlations exist (e.g., smoking correlating with cancer). In this case, the apparent attack success is driven by statistical inference rather than actual privacy violations, a phenomenon often referred to as a *privacy fallacy* [23, 43]. Indeed, several works establish that it is impossible to simultaneously provide utility and eliminate absolute information gain [23, 43].

We conclude that ReRo is unreliable as an attack resilience metric, as it overlooks key statistical phenomena that distort privacy risk assessment, such as data imputation and targeted attacks. Both cases are very common and have an impact in practice (see Section 7), motivating the need for a novel framework to more accurately assess the risk of DP mechanisms with respect to privacy attacks.

DP Auditing. DP auditing [3] seeks to demonstrate tight estimates of the privacy budget, discover implementation flaws, and estimate empirical privacy. However, auditing in practice remains a significant challenge. For instance, implementation bugs or design flaws can severely degrade privacy guarantees in ways that are not immediately obvious. To address this, black-box discovery methods such as DP-Sniper [13] and Eureka [48] have been developed to detect DP violations by training classifiers to distinguish between mechanism outputs from "worst-case" adjacent inputs. While effective at uncovering certain classes of violations, this assumption breaks down for frequency-oracle mechanisms over high-dimensional categorical domains, where outputs are discrete randomized encoding [7] with inherently combinatorial structure. Consequently, the learned classifiers fail to scale, becoming prohibitively slow or ineffective as the domain dimension grows.

Beyond identifying bugs, existing empirical privacy auditing approaches primarily focus on MIAs [4, 12, 37, 62], which limits their ability to detect broader forms of privacy leakage. Some auditing techniques extend beyond MIAs to consider AIAs, but these are restricted to specific contexts—such as Label DP [51] or synthetic data generation [35]. In the LDP setting, the state-of-the-art framework LDP Auditor [6] relies specifically on perfect reconstruction without target-specific auxiliary knowledge for auditing. Summarizing, despite its practical importance, no existing auditing framework supports a DRA-based analysis that goes beyond MIAs and enables systematic evaluation across diverse DP mechanisms.

4 RECONSTRUCTION ADVANTAGE

In this section, we introduce reconstruction advantage (RAD) as a novel, unifying metric for adversarial risk assessment. We first establish a worst-case bound on RAD that holds for any mechanism, data

distribution, and auxiliary knowledge, ensuring robustness when the attacker's prior knowledge is unknown. We then refine this result by deriving a tighter bound under known auxiliary knowledge and prove its tightness by constructing the corresponding optimal attack that achieves it. Together, these results provide a noise calibration method to optimize utility for a given risk. We empirically validate the practical tightness of our bounds in Section 7.3.

In order to address ReRo's lack of accounting for the impact of target-specific auxiliary knowledge, we explicitly incorporate this concept into RAD. Formally, each record $z \in \mathcal{Z}$ may be associated with target-specific auxiliary information $a(z) \in aux$. The auxiliary information can take different forms. For instance, in an AIA setting, where records are pairs $z = (x, y)$, one may define $a(z) = x$ and aim to infer y. Alternatively, in the image reconstruction setting, the target may be the full record z, while $a(z)$ could correspond to a label such as "image of a person" or "image of an animal". The only assumption we impose is that the type of auxiliary information is consistent across all records: if $a(z)$ corresponds to a set of pixels, then for any other record z', $a(z')$ must also be a set of pixels (and not, for example, a semantic label). Having established this formalization, we are now in a position to introduce our metric[1].

Definition 4.1 (η-RAD). Let π be a prior over $\mathcal{Z}$, $\ell\colon \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function, and $a(z) \in aux$ the target-specific auxiliary information for each $z \in \mathcal{Z}$. Given a mechanism $\mathcal{M} : \mathcal{Z}^n \to \mathcal{D}(\Theta)$, any dataset $D_- \in \mathcal{Z}^{n-1}$ and any adversary $A : \Theta \times aux \to \mathcal{D}(\mathcal{Z})$ we define the η-reconstruction advantage, η-RAD, as

$$\Pr_{\substack{Z_1\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_1})}}[\ell(Z_1, A(\theta, a(Z_1))) \leq \eta] - \Pr_{\substack{Z_0,Z_1\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_0})}}[\ell(Z_1, A(\theta, a(Z_1)) \leq \eta].$$

RAD explicitly accounts for target-specific auxiliary knowledge, providing a generalization of the membership and attribute advantages to arbitrary DRAs. Importantly, RAD takes values between -1 and $(1 - \kappa_\pi) \leq 1$ where $\kappa_\pi = \Pr_{Z,Z'\sim\pi}[Z = Z']$, i.e., the probability of resampling from the distribution π, analogously to membership and attribute advantage [70]. Intuitively, RAD measures the increase in the attacker's success probability that arises solely from the target's participation in the private learning process. In this way, RAD avoids the overestimation of risk that is inherent in ReRo. If RAD ≤ 0, participation carries no risk, since the attacker's probability of correctly reconstructing the record is no greater than if the individual had not participated. Larger values of RAD indicate higher participation risk. In the extreme case where RAD $= 1 - \kappa_\pi$, participation entails absolute risk: the attacker always succeeds in reconstructing the participant's record, while no sensitive information can be reconstructed from non-participants.

Previous bounds for ReRo assume that DRAs perform equally for every target. This assumption holds when the adversary has no target-specific auxiliary knowledge, but breaks once *aux* is available: for instance, knowing that a target's surname is "Smith" might give less information than knowing that it is "Sainthorpe-Burton", as the latter is less frequent and hence carries more information. Such differences are not captured by ReRo, nor reflected in the proofs of the corresponding bounds [9, 32], which consequently fail for attacks utilizing target-specific auxiliary knowledge as demonstrated in Section 7.3. Hence, we provide the first theoretical bound that explicitly accounts for *aux* and covers any possible attack from MIAs to the most general DRAs:

Theorem 4.2 ((ε, δ)-DP implies η-RAD). *Let $\pi, \ell, \eta \geq 0$ as in Def. 4.1, and $\kappa_\pi = \Pr_{Z,Z'\sim\pi}[Z = Z']$. If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies (ε, δ)-DP, then for any attack $A\colon \Theta \times aux \to \mathcal{D}(\mathcal{Z})$, and database D_- we have*

$$\eta\text{-RAD} \leq \mathrm{TV}(\mathcal{M})(1 - \kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1 - \kappa_\pi).$$

Proof Sketch. First, we prove that for every $z \in \mathcal{Z}$ and target-specific knowledge $a(z)$, any attack defines $A(D, a(z)) = \mathcal{A}_z(\mathcal{M}(D))$ verifying

$$\mathrm{TV}(\mathcal{A}_z(D_{z_1}), \mathcal{A}_z(D_{z_0})) \leq \mathrm{TV}(\mathcal{M}(D_{z_1}), \mathcal{M}(D_{z_0})). \tag{4}$$

Applying this property to the RAD definition, we obtain

$$\begin{aligned}
&\Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_0,Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)] \\
&= \mathop{\mathbb{E}}_{Z_0\sim\pi}\left[\mathop{\mathbb{E}}_{Z_1\sim\pi}[\Pr_{\mathcal{A}_{Z_1}}(S_\eta(Z_1) \mid D_{Z_1})] - \mathop{\mathbb{E}}_{Z_1\sim\pi}[\Pr_{\mathcal{A}_{Z_1}}(S_\eta(Z_1) \mid D_{Z_0})]\right] \\
&= \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\left(\Pr_{\mathcal{A}_{Z_1}}(S_\eta(Z_1) \mid D_{Z_1}) - \Pr_{\mathcal{A}_{Z_1}}(S_\eta(Z_1) \mid D_{Z_0})\right)\right] \\
&\overset{\text{Eq.4}}{\leq} \mathrm{TV}(\mathcal{M}) \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right].
\end{aligned}$$

Since, $\mathbb{E}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right] = 1 - \sum_z \pi_z^2$ for discrete variables and 1 for continuous ones, it follows the result. □

Note that in the discrete case, $\kappa_\pi = \sum_z \pi_z^2$, hence, the worst-case prior corresponds $\pi = U\{z_0, z_1\}$. In the continuous case, this result simplifies to η-RAD $\leq \mathrm{TV}(\mathcal{M})$, unaffected by the prior distribution.

Theorem 4.2 is the first bound for RAD under the strongest threat model, where the attacker may use auxiliary knowledge. Experiments on real datasets (see Section 7) show that this bound is tight: attacks can achieve the predicted advantage, confirming that it accurately captures the worst-case scenario. Hence, it is a crucial tool for DP noise calibration, improving over ReRo.

Moreover, Theorem 4.2 allows upper bounding RAD under composition. Given $\mathrm{TV}(\mathcal{M}_i) = \Delta$, the T-fold adaptative composition satisfies $\mathrm{TV}(M) \leq (1-(1-\Delta)^T)$. Hence, η-RAD$\leq (1-(1-\Delta)^T)(1-\kappa_\pi)$.

Since Theorem 4.2 does not depend on the attacker's auxiliary knowledge, the same bound holds whether the attacker has no auxiliary information ($aux = \{\varnothing\}$) or complete knowledge of the record ($a(z) = z$), since the result is derived in a worst-case manner. However, when the attacker's goal is to reconstruct an entire record (as in DRAs) or infer parts of it (as in AIAs), it is unreasonable to assume that the attacker already knows the full record ($a(z) = z$)— as assumed for MIAs. Therefore, we next provide a tighter bound that explicitly incorporates the target-specific auxiliary knowledge.

Theorem 4.3. *Given $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ and $a\colon \mathcal{Z} \to aux$ measurable, then for any attack $A\colon \Theta \times aux \to \mathcal{D}(\mathcal{Z})$, we have*[2]

$$\eta\text{-RAD} \leq \sum_{\theta\in\Theta}\sum_{x\in aux}\max_{z_\theta\in\mathcal{Z}}\sum_{\substack{\ell(z,z_\theta)\leq\eta \\ a(z)=x}} w(\theta, z)\pi_z$$

where $w(z, \theta) = p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)$.

[1]Note that we presented a preliminary idea for this metric in [32], initially calling it U-ReRo, which, however, similar to ReRo, fails to take *aux* into account.

[2]In the continuous case, the following sums must be changed by integrals and π_z by the density function (see long version for details).

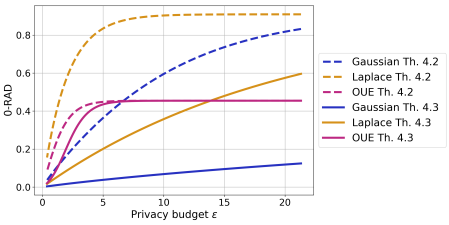


Figure 1: Improvement of Theorem 4.3 over Theorem 4.2 for different DP mechanisms, $|\mathcal{Z}| = 11$ and uniform prior.

Proof Sketch. We denote by μ the counting/Lebesgue measure and π mass/density function in the discrete/continuous case. Using probability properties, we rewrite RAD as

$$\int_{\mathcal{Z}} \pi_z \int_{\Theta} \Pr_A(S_\eta(z) \mid a(z), \theta) \underbrace{(p_{\mathcal{M}}(\theta \mid D_z) - p_{\mathcal{M}}(\theta))}_{w(z,\theta)} \, \mathrm{d}\mu(z) \, \mathrm{d}\mu(\theta).$$

For all z_1, z_2 such that $a(z_1) = a(z_2) = x$, and for any θ,

$$\Pr_A(S_\eta(z_1) \mid a(z_1), \theta) = \Pr_A(S_\eta(z_1) \mid a(z_2), \theta) = \Pr_A(S_\eta(z_1) \mid x, \theta).$$

Hence, given $a^{-1}(x) = \{z \colon a(z) = x\}$ for all $x \in aux$, given $\nu = \mu \circ a^{-1}$, applying the disintegration theorem [8] it exists a unique measure μ_x such that

$$\begin{aligned}
\eta\text{-RAD} &= \int_{\mathcal{Z}} \int_{\Theta} \Pr_A(S_\eta(z) \mid a(z), \theta) w(z,\theta) \pi_z \, \mathrm{d}\mu(z) \, \mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(x)} \Pr_A(S_\eta(z) \mid x, \theta) w(z,\theta) \pi_z \, \mathrm{d}\mu_x(z) \, \mathrm{d}\nu(x) \, \mathrm{d}\mu(\theta) \\
&\leq \int_{\Theta} \int_{aux} \max_{z_\theta \in Z} \int_{S_\eta^x(z_\theta)} w(z,\theta) \pi_z \, \mathrm{d}\mu_x(z) \, \mathrm{d}\nu(x) \, \mathrm{d}\mu(\theta)
\end{aligned}$$

where $S_\eta^x(\tilde{z}) = \{z \colon a(z) = z \wedge \ell(z, \tilde{z}) \leq \eta\}$. □

Theorem 4.3 bounds RAD when the specific $\mathcal{M}$ and auxiliary knowledge, *aux*, are known. At the same time, it becomes more precise than our worst-case bound Theorem 4.2. Moreover, it admits simple characterizations for commonly studied threat models. Particularly, if in a MIA, i.e., $a(z) = z$ for all records, we get

$$\eta\text{-RAD} \leq \sum_{z \in \mathcal{Z}} \sum_{\theta \colon w(\theta,z) > 0} w(\theta, z) \pi_z, \tag{5}$$

since $\arg\max_{z_\theta} = S_\eta(z)$ if $w(\theta, z) > 0$ and $\arg\max_{z_\theta} = \mathcal{Z} \setminus S_\eta(z)$ otherwise. On the other extreme, when $aux = \{\varnothing\}$,

$$\eta\text{-RAD} \leq \sum_{\theta \in \Theta} \max_{z' \in \mathcal{Z}} \sum_{\ell(z', z) \leq \eta} w(\theta, z) \pi_z. \tag{6}$$

Moreover, if $\eta = 0$ (perfect reconstruction), as in AIA and the original ReRo setting [34]), Theorem 4.3 equation simplifies to :

$$0\text{-RAD} \leq \sum_{\theta \in \Theta} \sum_{x \in aux} \max_{\substack{a(z) = x \\ w(z,\theta) > 0}} w(z, \theta) \pi_z, \tag{7}$$

We illustrate the benefits of Theorem 4.3 on relevant DP mechanisms through next examples and visualizations in Figures 1 and 4.

Example 4.4. The generalized randomized response mechanism (GRR) [41] is an LDP mechanism that outputs the true record z_1 with probability $p = \mathrm{e}^\varepsilon / (\mathrm{e}^\varepsilon + m - 1)$ and any other record $z_0 \neq z_1$ with probability $q = (\mathrm{e}^\varepsilon + m - 1)^{-1}$. Since, $p \geq q$ for all $\varepsilon \geq 0$,

$$w(\theta, z) = \begin{cases} (p - q)(1 - \pi_\theta) & \text{if } z = \theta \\ (q - p)\pi_\theta & \text{otherwise,} \end{cases} \tag{8}$$

and $w(z, \theta) > 0$ iff $z = \theta$. Hence, applying Theorem 4.3 for $a(z) = z$:

$$\eta\text{-RAD} = \sum_{\theta} (p - q)(1 - \pi_\theta)\pi_\theta = \tfrac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + m - 1}(1 - \kappa_\pi) = \mathrm{TV}(1 - \kappa_\pi).$$

While, if we consider $aux = \{\varnothing\}$,

$$\eta\text{-RAD} = (p - q)(1 - \sum_{\theta} \pi_\theta \inf_{\ell(z_\theta, \theta) \leq \eta} \Pr_{Z \sim \pi}[\ell(Z, z_\theta) \leq \eta]).$$

Example 4.5. The optimal unary encoding (OUE) mechanism [66] maps each input $z \in \mathcal{Z}$ to an m-dimensional one-hot binary vector and perturbs each bit independently. For each position $i \in [m]$, the obfuscated vector θ is sampled such that $\Pr[\theta_i = 1] = 1/2$ if $i = z$, and $\Pr[\theta_i = 1] = q = \frac{1}{\mathrm{e}^\varepsilon + 1}$ otherwise. Denoting $p = 1 - q$, according to Theorem 4.3, we obtain that, for $a(z) = z$:

$$\eta\text{-RAD} \leq \frac{1}{2} \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1}(1 - \kappa_\pi) = \mathrm{TV}(\mathrm{OUE})(1 - \kappa_\pi).$$

If we consider $aux = \{\varnothing\}$, then the bound becomes:

$$0\text{-RAD} \leq \frac{p - q}{2p} \left(\sum_{i=1}^{m} p^{m-i} \pi_i (1 - \pi_i) - q \sum_{i=1}^{m} p^{m-i} \pi_i \sum_{z=1}^{i-1} \pi_z \right)$$

which in particular for $\pi = U[m]$:

$$0\text{-RAD} \leq \frac{(2p - 1)\left(1 - p^{m-1}\right)}{2m(1 - p)} = \frac{\mathrm{e}^\varepsilon - 1}{2m} \left(1 - \left(\frac{\mathrm{e}^\varepsilon}{1 + \mathrm{e}^\varepsilon} \right)^{(m-1)} \right).$$

Note that when $\varepsilon \to \infty$ previous bound converges to $\frac{m-1}{2m}$, hence even if we keep reducing the noise (increasing ε), the attacker's advantage is limited.

Example 4.6. In the subset selection mechanism (SS) [69] users report a subset $\theta \subseteq \mathcal{Z} = \{z_1, \ldots, z_m\}$ containing their true value z with probability $p = \frac{\omega \mathrm{e}^\varepsilon}{\omega \mathrm{e}^\varepsilon + m - \omega}$, where $\omega = |\theta| = \max\left(1, \left\lfloor \frac{m}{\mathrm{e}^\varepsilon + 1} \right\rfloor\right)$. The subset is completed by sampling uniformly from $\mathcal{Z} \setminus \{z\}$. According to Theorem 4.3 we obtain that for $\pi = U[m]$

$$0\text{-RAD} \leq \frac{pm - \omega}{m\omega}.$$

Example 4.7. The Laplace mechanism adds Laplace noise with scale $b = \Delta q / \varepsilon$ to the query value $q(D) \in \mathbb{R}$ [25]. If $\mathcal{Z} = \{z_1, \ldots, z_m\}$ is uniformly distributed and $\Delta q = 1$ applying Theorem 4.3 we obtain

$$0\text{-RAD} \leq \frac{m - 1}{m} \left(1 - \mathrm{e}^{-\frac{\varepsilon}{2(m-1)}} \right).$$

Example 4.8. The Gaussian mechanism adds Gaussian noise $\mathcal{N}(0, \sigma)$ the query value $q(D) \in \mathbb{R}$ [10]. Given Φ the CDF of the standard normal distribution, if $\mathcal{Z} = \{z_1, \ldots, z_m\}$ is uniformly distributed and $\Delta q = 1$, applying Theorem 4.3 we obtain

$$0\text{-RAD} \leq \frac{m - 1}{m} \left(2\Phi\left(\frac{1}{2\sigma(m - 1)} \right) - 1 \right).$$

These examples demonstrate the applicability of Theorem 4.3 to estimate the risk in real-world scenarios. In Figure 1 we see the improvement when we target specific auxiliary knowledge instead of using our worst-case bound (Theorem 4.2). Hence, Theorem 4.3

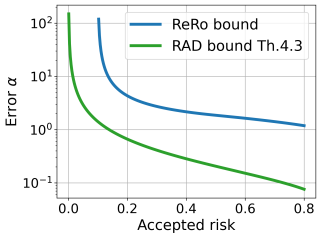


Figure 2: Upper bound on the Laplace mechanism query error (utility) at 95% confidence when the noise is calibrated using ReRo vs. RAD. We see that for the same risk estimation, calibrating with using RAD improves utility.

offers an improved noise calibration method to ensure protection against real attacks, when the auxiliary knowledge is well defined. For instance, when the entire record is considered private ($aux = \{\varnothing\}$); alternatively, when a specific attribute y is deemed sensitive, and we consider all the remainder record public, i.e., $a(z) = z \backslash y$.

Importantly, we illustrate in Figure 2 the utility gain of noise calibration using our RAD bounds compared to using the best existing ReRo bound [34], showing the benefit of our bounds for system design. Specifically, we consider $aux = \{\varnothing\}$—allowing comparison with [34]. We plot the upper bound on the Laplace mechanism's query error that can be guaranteed with 95% confidence, for $|\mathcal{Z}| = 10$ and $\Delta = 1$, showing a substantial improvement in utility enabled by our RAD-based calibration.

Crucially, Theorem 4.3 is universally tight: for any $\mathcal{M}$ and aux, there is an attack achieving the bound, so it cannot be further improved. We establish this result by explicitly constructing such attack in Algorithm 1. This result is particularly relevant to the database community, as it shows that, for any accepted risk tolerance, the utility achievable by $\mathcal{M}$ is fundamentally limited by our bound. In other words, our method yields an optimal noise calibration.

Corollary 4.9 (Attack Optimality). *Given the conditions as in Theorem 4.3, Algorithm 1 achieves the highest attainable η-RAD.*

Corollary 4.9 directly establishes that Theorem 4.3 is universally tight and Theorem 4.2 is tight, since there exists at least one mechanism (GRR Example 4.4) for which the bound of Theorem 4.2 is achieved. We further validate that this is not an isolated case by empirically demonstrating tightness on additional mechanisms, such as DP-SGD (See Figure 3c).

Beyond the theoretical contribution, our results provide a practical tool: a general attack algorithm that practitioners can directly use to evaluate the privacy risks of their systems or the tightness of their bounds. As a concrete demonstration, we apply this attack in the context of LDP auditing (see Section 6) and to assess empirical risk and tightness of our bounds in Section 7.

Our bounds offer concrete guidance for algorithm design, as they can be directly leveraged for noise calibration to achieve rigorous privacy guarantees while maximizing utility. In particular, they induce a simple protocol for practitioners. First, one must specify which information is deemed private (e.g., the full record, a subset of attributes, or membership), which determines the choice of the auxiliary information aux and $a\colon \mathcal{Z} \to aux$. Second, if prior

Algorithm 1: Optimal Attack.

Input : θ and $a(z) = x$
Output: $\tilde{z}$
Compute $a^{-1}(x) = \{z\colon a(z) = x\}$
for $z' \in \mathcal{Z}$ **do**
  $\mathcal{W}_\eta^x(z') = \sum_{z \in a^{-1}(x)\colon \ell(z,z') \le \eta} w(\theta, z)\pi_z$;
Select $\tilde{z} \in \arg\max_{z'} \mathcal{W}_\eta^x(z')$ (at random)

knowledge about the distribution of $\mathcal{Z}$ is available, it should be encoded in a distribution π. If this is not the case, however, one must resort to the worst-case prior; otherwise, the attacker's risk may be underestimated. This worst-case prior typically corresponds to $\pi_x = \pi_y = 1/2$ for the two records that are easiest to distinguish (see Examples 4.4 and 4.5 and Figure 4). Nevertheless, even when the worst-case prior cannot be explicitly identified, the total variation bound given in Theorem 4.2 provides a safe upper bound for any choice of prior and auxiliary knowledge.

Third, the resulting RAD of the mechanism can be computed using Theorem 4.3—an auxiliary-dependent bound proven to be universally tight, or upper-bounded by a worst-case guarantee when the nature of aux is unknown (Theorem 4.2). Finally, by inverting the corresponding bound, one can directly derive the noise-injection parameters that meet a prescribed risk level. Since our bounds are tight, this procedure yields mechanisms that are utility-optimal for any given risk acceptance.

Note that while the closed form of Theorem 4.3 is easy to derive for discrete data, this may not hold for continuous data, where the bound involves Lebesgue integrals. In such case, the bound can be evaluated numerically using a nested Monte Carlo procedure. While numerical approximations introduce error, we show in the long version of this paper how to obtain controlled confidence intervals in practice. As a safer alternative, one may always use our closed-form upper bound in Theorem 4.2. However, this bound can be overly conservative when $aux = \{\varnothing\}$, motivating the tighter closed-form upper-bounds derived in the next section, which avoid numerical procedures even for continuous data.

5 η-RAD UPPER BOUNDS UNDER $aux = \{\varnothing\}$

Our bound in Theorem 4.3 is universally tight, but two limitations remain. First, it requires full knowledge of the mechanism, making it suitable for noise calibration; however, in DP auditing, we often have only query access (e.g., auditing external software) without insight into the internal protocol [31]. Second, the bound lacks a closed form hence may rely on numerical approximation, particularly for continuous data domains. Consequently, in this section we provide black-box bounds for the case $aux = \{\varnothing\}$, both because this is the standard assumption in prior DP auditing [6, 50] and data reconstruction studies [9, 34], and because it makes practical sense: for other auxiliary-information models, one can always rely on the closed-form bound provided by Theorem 4.2.

First, we present a general bound that applies to any reconstruction setting as long as no target-specific auxiliary knowledge is available. For this purpose, we introduce $\kappa^-_{\pi,\ell}(\eta)$ as the infimum

counterpart of $\kappa^{+}_{\pi,\ell}(\eta)$, formally defined as

$$\kappa^{-}_{\pi,\ell}(\eta) = \inf_{z_0 \in \mathcal{Z}} \Pr_{Z\sim\pi}\left[\ell(Z, z_0) \le \eta\right], \tag{9}$$

representing the success probability of an oblivious attacker attempting to reconstruct the most difficult target only using π.

Theorem 5.1. *If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies f-DP, then for any attack with $aux = \{\varnothing\}$, $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, it satisfies*

$$\eta\text{-RAD} \le \max_{\alpha\in[\kappa^{-}_{\pi,\ell}(\eta),\kappa^{+}_{\pi,\ell}(\eta)]} 1 - f(\alpha) - \alpha.$$

If $\mathcal{Z}$ is discrete it also holds

$$\eta\text{-RAD} \le (1-\kappa_\pi) \max_{\alpha\in[0,\frac{\kappa^{+}_{\pi,\ell}(\eta)}{1-\kappa_\pi}]} 1 - f(\alpha) - \alpha.$$

Proof Sketch. We denote $\mathcal{A} = A \circ \mathcal{M}$ and combine Definition 2.4 and f convexity, to obtain the following upper bound

$$1 - f\Big(\mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\big[\Pr_{\mathcal{A}}[S_\eta(Z_1) \mid D_{Z_0}]\big]\Big) - \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\big[\Pr_{\mathcal{A}}[S_\eta(Z_1) \mid D_{Z_0}]\big]$$

and we prove that under $aux = \{\varnothing\}$ assumption,

$$\kappa^{-}_{\pi,\ell}(\eta) \le \big[\Pr_{\mathcal{A}}[S_\eta(Z_1) \mid D_{Z_0}]\big] \le \kappa^{+}_{\pi,\ell}(\eta),$$

finishing the result. For the discrete case, we show that RAD $\le (1-\kappa_\pi)\,(1 - f(B) - B)$ with

$$B = \tfrac{1}{1-\kappa_\pi} \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}} \Pr_{\mathcal{A}}[S_\eta(Z_1) \mid D_{Z_0}]\right] \in [0, \tfrac{\kappa^{+}_{\pi,\ell}(\eta)}{1-\kappa_\pi}]. \quad \square$$

This result serves as an upper-bound approximation of RAD, when $aux = \{\varnothing\}$. Moreover, as a consequence of the previous result, we obtain a general result for any (ε,δ)-DP mechanism:

Proposition 5.2. *If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies (ε,δ)-DP, then for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, it satisfies*

$$\eta\text{-RAD} \le \min\{\kappa^{+}_{\pi,\eta}(e^\varepsilon - 1) + \delta, \tfrac{(1-\kappa^{-}_{\pi,\eta})(e^\varepsilon-1)+\delta}{e^\varepsilon}, \tfrac{e^\varepsilon-1+2\delta}{e^\varepsilon+1}(1-\kappa_\pi)\}.$$

Proof Sketch. The claim follows by combining Theorem 5.1 with the observation that any (ε,δ)-DP mechanism is f-DP with $f(\alpha) = \max\{0, 1-\delta-e^\varepsilon\alpha, \tfrac{1-\delta-\alpha}{e^\varepsilon}\}$ [22], and maximize all cases. $\square$

Next, we focus on perfect reconstruction, i.e., $\eta = 0$, in categorical data. This case is particularly relevant since many sensitive attributes, such as diseases, political opinions, or religious beliefs, are categorical and do not trivially support partial reconstruction, e.g. [28, 29]. For such settings, we derive more precise bounds:

Theorem 5.3 (0-RAD under (ε,δ)-DP). *Given $|\mathcal{Z}| = m$ with prior $\pi_1(1-\pi_1) \ge \cdots \ge \cdots \ge \pi_m(1-\pi_m)$ and $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ an (ε,δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$*

$$0\text{-RAD} \le \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1} K_\pi + R \max_{i>K} \pi_i$$

where $K \in [m]$ is the largest index satisfying $R = (m-1)\tfrac{e^\varepsilon-1+m\delta}{e^\varepsilon+m-1} - (K - \sum_{i=1}^{K}\pi_i)\tfrac{e^\varepsilon-1+2\delta}{e^\varepsilon+1} \ge 0$ and $K_\pi = \sum_i^K (1-\pi_i)\pi_i$.

Proof Sketch. Given $|\mathcal{Z}| = m$ and $aux = \{\varnothing\}$, we rewrite Theorem 4.2

$$0\text{-RAD} \le \sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i)\right)\pi_z, \tag{10}$$

where $\Theta_1 = \{\theta \in \Theta\colon z_1 \in \arg\max_j w(\theta, z_j)\pi_j\}$ and for every $i \ge 1$, Θ_{i+1} is recursively defined as $\Theta_{i+1} = \{\theta \in \Theta\colon z_{i+1} \in \arg\max_j w(\theta, z_j)\pi_j\} \setminus \cup_{k=1}^{i}\Theta_k$. Then, using DP properties and Θ_z definition we prove that

$$\Gamma = \sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i)\right) \le \frac{(m-1)(e^\varepsilon - 1 + \delta m)}{e^\varepsilon + m - 1}. \tag{11}$$

Finally, we prove that $\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) \le \mathrm{TV}(\mathcal{M})(1-\pi_i)$ and combine it with Equation (11) obtaining the result. $\square$

Note that in the extreme case where $\pi_1 = \pi_2 = \frac{1}{2}$ and $\pi_i = 0$ for all $i \neq 1, 2$, we recover exactly the same result as in Theorem 4.2. This formulation enables the assessment of intermediate configurations of π. Notably, when $\pi = U[m]$ yields a marked improvement:

Corollary 5.4 (Black-box Uniform Prior). *Given $\pi = \mathrm{U}[m]$ the uniform distribution over $\mathcal{Z}$. If a mechanism $\mathcal{M}$ satisfies (ε,δ)-DP, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$ it guarantees*

$$0\text{-RAD} \le \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}.$$

Remark on Composition. Since our η-RAD bounds depend explicitly on the privacy parameters—namely ε, δ, and/or f—they can be directly recomputed under composition by first applying the corresponding composition results to obtain the composed privacy parameters (Cf. Section 2), and then evaluating the bounds on these composed values. In the following example, we illustrate how to derive RAD composition bounds for the particular case of DP-SGD.

Example 5.5. Given a risk threshold, RAD $\le \gamma$, we aim to calibrate the noise scale σ on a full-batch DP-SGD (i.e., the standard deviation of the Gaussian noise added to the gradients during training [1]), for T steps to protect against the threat model considered by Hayes et al. [34], i.e., white-box access to private gradients, uniform prior over $|\mathcal{Z}| = m$ and $\eta = 0$, hence $\kappa_- = \kappa_+ = 1/m$.

Each iteration of a full-batch DP-SGD is (σ^{-1})-GDP [22], hence by f-DP composition rule, T iterations of DP-SGD are $(\sqrt{T}\sigma^{-1})$-GDP (cf. Section 2). Combining this composition result with our theorems we obtain direct calibration rules:

Without information about *aux*, we use Theorem 4.2. Any μ-GDP mechanism has total variation TV $\le 2\Phi(\frac{\mu}{2}) - 1$ [30], hence DP-SGD after T iterations satisfies $\gamma \le \frac{m-1}{m}(2\Phi(\frac{\sqrt{T}}{2\sigma}) - 1)$. We plot this bound for $T = 100$ in Figures 3b and 3c.

If we consider the whole records sensitive, $aux = \{\varnothing\}$, then we apply Theorem 5.1:

$$0\text{-RAD} \le \frac{m-1}{m} \max_{\alpha\in[0,\frac{1}{m-1}]}\left(1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{T}}{\sigma}\right) - \alpha\right)$$

Hence, given $\alpha^* = \min\left\{\frac{1}{m-1},\ 1 - \Phi\left(\frac{\sqrt{T}}{2\sigma}\right)\right\}$, the minimum σ to guarantee 0-RAD $\le \gamma$ is:

$$\sigma \ \ge\ \frac{\sqrt{T}}{\Phi^{-1}(1-\alpha^*) - \Phi^{-1}\left(1 - \frac{m}{m-1}\gamma - \alpha^*\right)}.$$

Using this bound a practitioner can choose the minimum noise scale σ for a risk threshold. For example, if $m = 10$ and individuals accept risk $\gamma \le 0.1$, then $T = 100$ training iterations require noise $\sigma = 22$. We plot this bound for the case of $T = 100$ in Figure 3a.

In summary, this section provides reasonable closed-form upper bounds (as we show in Section 7.3) for estimating RAD when Theorem 4.3 cannot be computed explicitly or $\mathcal{M}$ is unknown and $aux = \{\varnothing\}$, hence Theorem 4.2 would overestimate the risk. Importantly, these bounds offer composition results.

6 RAD FOR DP AUDITING

DP auditing is crucial for assessing the tightness of DP mechanisms, establishing the practical impact of the mechanism parameters, and detecting implementation flaws in deployed DP mechanisms [3, 13, 37]. While previous DP auditing tools focus on solving specifically one of the aforementioned aspects, we propose a general-purpose DP auditing framework: RAD-based DP auditing.

RAD provides a unifying framework for analyzing adversarial risk under arbitrary threat models. Moreover, our bounds establish a tight and explicit connection between RAD and the standard privacy parameters. Taken together, these results yield a simple and principled approach to general-purpose DP auditing. Precision and tightness are especially critical in this context, since loose estimates may underestimate privacy risks or fail to detect bugs and implementation flaws.

The core idea of RAD-based auditing is straightforward: given a measured RAD value $\tilde{\gamma}$, we invert our theoretical bounds to estimate an empirical privacy budget. This empirical $\tilde{\varepsilon}$ reflects the observed privacy loss in practice, complementing theoretical worst-case values and providing a more realistic perspective on real-world risk. Formally, in previous sections, we provide bounding functions B such that $\mathrm{RAD}(\mathcal{M}) \leq B(\varepsilon, \delta)$ for any (ε, δ)-DP mechanism. Given a bound η-RAD$\leq B(\varepsilon, \delta)$, we compute RAD empirically obtaining γ, and estimate $\tilde{\varepsilon} \geq B^{-1}(\gamma, \delta)$.

The bound we employ depends on the specific setting. For instance, in a completely black-box scenario—where not even the mechanism used is known—for categorical data, in which we assume $\pi = U[m]$, the best bound is Corollary 5.4. Therefore, the DP auditing framework consists of running an attack, measuring its empirical RAD $\widetilde{\gamma}$, and deriving $\widetilde{\varepsilon}$ as follows:

$$\widetilde{\varepsilon} = \begin{cases} \ln\left(\frac{\widetilde{\gamma}\, m+1}{1-\widetilde{\gamma}\frac{m}{m-1}}\right) & \text{if the term can be evaluated,} \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{12}$$

However, if the mechanism $\mathcal{M}$ is known, we can use our improved bound from Theorem 4.3 (See examples 4.4 to 4.6).

Our auditing framework overcomes the fundamental scalability limitations of prior learning-based approaches such as DP-Sniper and Eureka [13, 48], enabling auditing in high-dimensional categorical LDP settings. Unlike these methods, our approach avoids costly hyperparameter tuning and the search for worst-case neighboring databases, and remains computationally feasible even when the input domain contains thousands of categories (see Section 7).

Despite the importance of LDP mechanisms [27], only one major work has so far focused on LDP auditing: LDP Auditor [6]. Applying our RAD-based DP auditing to LDP, we address key limitations of prior work. In contrast to LDP Auditor, which focuses exclusively on perfect reconstruction without target-specific auxiliary knowledge—excluding important use-cases such as AIAs—we allow auditing under broader threat models by leveraging optimal attacks (see Algorithm 1). Moreover, our approach is not constrained by internal parameter choices that bound the maximum privacy loss estimate (as in LDP Auditor) [6], thus providing tighter and more accurate guarantees. We investigate and empirically show the improvement in accuracy of our auditing approach in Section 7 (cf. Figure 5 for results), where we audit three main LDP mechanisms—GRR, SS and OUE—showing improved accuracy for all of them.

7 EXPERIMENTS

In this section, we empirically examine the limitations of ReRo described in Section 3, focusing on how existing bounds fail to account for realistic attackers with target-specific auxiliary information. Moreover, we validate our theoretical bounds and our RAD-based DP auditing framework in real-world databases and DP mechanisms. Our experiments show that RAD accurately distinguishes privacy leakage from imputation, with tight bounds in practice, making it a reliable tool for interpretable noise calibration. RAD also enables auditing of LDP mechanisms, improving both scope and accuracy over the state-of-the-art [6].

7.1 Database Description

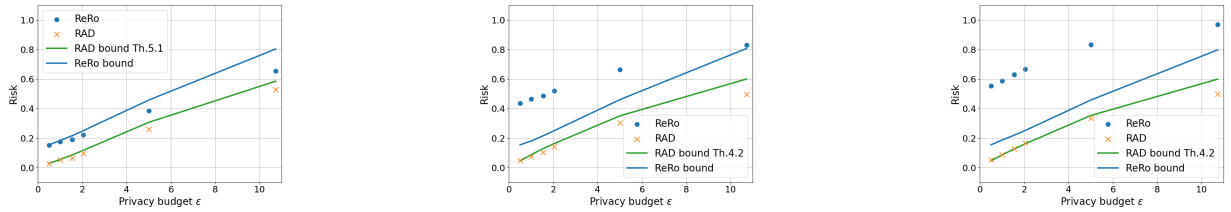
We evaluate private learning, aggregation and LDP scenarios, using tailored datasets for each setting. The database selection is guided by their relevance in prior work and availability.

For DP-SGD, we use the same dataset as in ReRo [34] for consistency: MNIST [45], with 70,000 grayscale images of handwritten digits. We also replicate results on Fashion-MNIST [67] (Fashion), which similarly contains 70,000 grayscale images of clothing items.

To evaluate the imputation attack [40], we use the Census and Texas-100X datasets in consistency with the original paper. The Census dataset [40] contains 1,676,013 records with 14 attributes, where race is treated as the sensitive attribute with eight categories. The Texas-100X dataset [40] comprises 925,128 patient records from 441 hospitals, including demographic and medical attributes, with a binary ethnicity attribute designated to be sensitive.

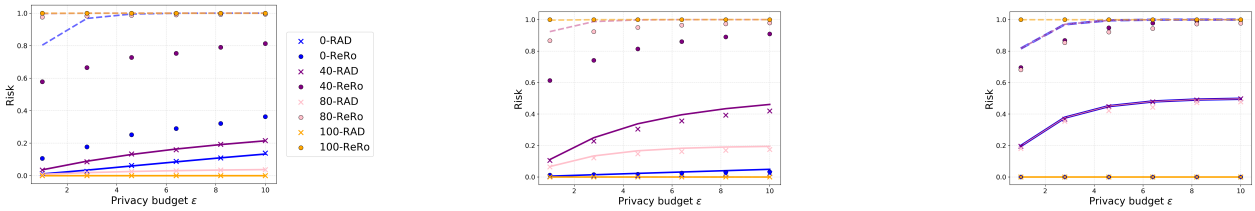
We evaluate aggregation in the Adult dataset [11], a census dataset commonly used in privacy-preserving aggregation [61]. It consists of 32, 561, records with two numerical attributes, from which we select (working) hours-per-week following previous work [61], leading to the domain $\mathcal{Z} = \{0, \ldots, 100\}$.

We evaluate our LDP auditing framework on reconstruction attacks against location data using two real-world mobility datasets: the Porto dataset [57] and the Geolife dataset [71]. Both datasets are widely used in privacy and mobility research (e.g., [47, 59, 68]) and are publicly available. Each dataset consists of GPS coordinates, which we map to the OpenStreetMap (OSM) graph format [58] like prior work. The Porto dataset contains a total of 83, 409, 386 location reports that we map to the OSM roadgraph at Porto's city center (41.1475, -8.5870) with a 2.7 km radius, capturing the urban core of Porto. This radius leads to a universe size $|\mathcal{Z}| = 3,052$. The Geolife dataset contains a total of 24, 876, 978 locations that we mapped to an OSM graph centered near Tiananmen Square (39.9130, 116.3703) with a 5 km radius covering major central districts, leading to a universe of size $|\mathcal{Z}| = 5,356$.



(a) DRA, $aux = \{\varnothing\}$. **(b) DRA, $a(z)$ = image label.** **(c) MIA: $a(z) = z$.**

Figure 3: RAD vs ReRo results for optimal attacks against DP-SGD on MNIST. Lines show theoretical bounds and markers of empirical risk as estimated by RAD/ReRo. Empirical results exceed the bounds as estimated by ReRo, RAD bounds hold.



(a) Adult distribution. **(b) Uniform distribution.** **(c) $\pi(0) = \pi(100) = \frac{1}{2}$.**

Figure 4: RAD vs ReRo results for optimal attack against Truncated Laplace on Adult. Straight lines show RAD bounds (Theorem 4.3) and dashed lines ReRo bounds ([34]). Markers show empirical risk as estimated by RAD/ReRo.

7.2 Experiment Design

We investigate attacks on private learning (DP-SGD), aggregation queries (Laplace mechanism), and LDP protocols (GRR, OUE, SS) under varying auxiliary information settings to validate our bounds, compare RAD and ReRo, and evaluate our auditing framework.

We demonstrate *ReRo overestimating risk* due to imputation and how RAD overcomes this with the pure imputation attack [40]: It uses a public dataset D_- to train a separate attack classifier A_I that, given the public attributes of a target, returns as label a prediction for the sensitive one. The adversary is given only the target public attribute $a(z)$ and outputs the prediction $\widetilde{s_z} = \arg\max_{s_i \in \Theta} \Pr_I[s_i \mid a(z)]$, where the conditional distribution $\Pr[s_i \mid a(z)]$ is estimated by A_I, once the imputation model has been trained on D_-. This attack does not use any information from the target model $\mathcal{M}(D)$; therefore, adversarial success cannot be privacy leakage resulting from a user's participation in the training dataset of $\mathcal{M}(D)$. Following the original paper [40], we tested in both the Census and Texas datasets. We set $|D_-| = 49,000$ and a universe $\mathcal{Z}$ of $m = 1000$, randomly selected from the remaining data records consistent with [40]. We define the attack to be successful, $\ell(z, z') = 0$, if $a(z) = a(z')$, as is typical for AIAs.

We show *RAD improvement over ReRo* and the optimality of our bounds both in private learning and DP aggregation. In both cases, we test our optimal attacks to assess tightness.

For private learning, we run the attacks against DP-SDG on the MNIST and Fashion image datasets in three settings: $aux = z$ (a MIA), $aux = \{\varnothing\}$ (a DRA, replicating the setting in [34]), and $aux = a(z)$ (a DRA, where the adversary also knows the target image's label, i.e., which object is contained). We declare an attack successful when $A(\theta, a(z)) = z$, that is, $\eta = 0$. We set $|D_-| = 999$ (and so the training set size is $|D_- \cup \{z\}| = 1,000$) and train with full-batch DP-SGD for $T = 100$ steps. We set the clipping rate, i.e., the maximum norm we clip the real gradients to while training, $C = 0.1$ and $\delta = 10^{-5}$ and adjust the noise scale σ (see Example 5.5) for a given target ε. We set the uniform prior with size $|\mathcal{Z}| = 8$ (disjoint from D_-), meaning that $\kappa^+_{\pi,0} = \kappa_\pi = 0.125$. Hence, we exactly replicate the original ReRo study [34] parameters.

For DP aggregation, we evaluate the optimal attack against the Laplace mechanism on sum queries using the "working-hours" attribute of Adult, employing truncation as a post-processing operation. We empirically compute the distribution π from the original data to simulate a real-world setting (reflecting that working 40 h/week is apriori more likely than working 100 h/week), a uniform distribution, and a completely skewed distribution with $\pi(100) = \pi(0) = 1/2$. For all cases, we set $|D| = 999$, $aux = \{\varnothing\}$ and evaluate the performance for $\eta \in \{0, 40, 80, 100\}$.

Finally, *we evaluate RAD in LDP*, and we compare our auditing framework with the state-of-the-art tool LDP Auditor [6] for three relevant LDP mechanisms: GRR, OUE and SS [7, 33]. To obtain the results for LDP Auditor, we used the code from Arcolezi and Gambs's public GitHub repository [5]. LDP Auditor estimates the empirical privacy budget in 10^6 runs.

We evaluate RAD based on our optimal attack (See Alg. 1) under a uniform prior and without auxiliary knowledge, allowing comparison with LDP Auditor. We then test our own LDP auditing framework: based on the obtained RAD value γ, we evaluate $B^{-1}(\gamma)$ for B following Theorem 4.3 and obtain an estimate of the empirical privacy budget. The precise $B(\varepsilon)$ for GRR, OUE and SS are shown in Examples 4.4 to 4.6 respectively. Since B^{-1} is not explicit for OUE, we approximate it numerically using the bisection method, which

Dataset	ReRo	RAD
Census	0.81	0
Texas	0.73	0

Table 1: ReRo Vs. RAD risk estimation for imputation attack.

converges in $O(\log(\tau^{-1}))$ iterations, where τ denotes the tolerance level [60]. We set $\tau = 10^{-6}$. Consistent with [6], we repeat the ε estimation five times and report the mean and standard deviation.

All experiments rely on empirical estimates of ReRo and RAD. To obtain these estimates, we use Monte Carlo methods, approximating expected values by repeatedly sampling from the random process and computing the average. Following [34], ReRo is estimated by repeating J times the attack $A(\mathcal{M}(D_z), a(z))$ for each $z \in \mathcal{Z}$ and computing the π-weighted average. The RAD correction term is estimated analogously by evaluating J times the attacks $A(\mathcal{M}(D_{z_0}), a(z_1))$ for each target–challenger pair $z_1, z_0 \in \mathcal{Z}$ and averaging the results.

For MNIST, Fashion and Adult, we set $J = 1,000$ (as in [34]). Note that in the LDP cases $D_- = \varnothing$, and we set $J = 10^6/m$ ensuring the total number of runs matches those 10^6 repetitions of LDP Auditor. Finally, for the imputation attack, we do not require a target model as it is target model-independent and set $J = 1$. We repeat the imputation attack with five different seeds and report the averaged ReRo and RAD scores.

We use Python and TensorFlow [64] to evaluate the attacks. For DP-SGD ReRo, we rely on a minimal implementation provided by Hayes et al. [34], which we extend to incorporate RAD and target-specific auxiliary knowledge. For the imputation attack [40], we adapt the authors' public implementation [39].

7.3 Results

In this section, we present the results of RAD and ReRo empirical risk estimates and their corresponding theoretical bounds. For both ReRo and RAD, the y-axis shows the risk measure, with values near one indicating high risk and near zero indicating low risk.

7.3.1 RAD covers, but ReRo breaks for auxiliary knowledge. Figure 3 shows the results of ReRo and RAD risk estimation for our optimal attacks against DP-SGD on the MNIST dataset. Analogous results for the Fashion dataset are provided in the long version of the paper. We also include the corresponding theoretical bounds for ReRo and RAD for comparison. As expected, the existing bounds for ReRo [34] correctly upper-limit the empirically observed ReRo risk when the adversary has no prior knowledge of the victim record ($aux = \{\varnothing\}$, Figure 3a). However, when the adversary has prior knowledge of the victim record (Figures 3b and 3c), ReRo reveals higher disclosure than predicted by its theoretical bounds. In contrast, our RAD bounds consistently upper-limit the empirically estimated RAD risks across all tested attacks.

This supports our expectation that the ReRo bound only holds under the assumption that the adversary has no auxiliary knowledge about the victim ($aux = \{\varnothing\}$), but fails to correctly estimate privacy risks when target-specific auxiliary knowledge exists.

We can also observe that our bounds for RAD overcome this estimation error: they hold for any auxiliary knowledge and are nearly tight. In particular, Figures 3b and 3c show that the tightness of our worst-case bound Theorem 4.2 is not an isolated feature of GRR, but a reliable property that also applies to other widely used mechanisms, such as DP-SGD. Finally, Figure 3a shows that our closed-form bound Theorem 5.1 offers a reasonable upper-bound also when Theorem 4.3 needs to be numerically approximated (as is the case, for instance, with DP-SGD).

7.3.2 Leakage vs. Imputation. Table 1 compares the risk estimates of RAD and ReRo for the imputation attack. This attack is not based on any information leakage from the mechanism and ignores any output in the process. RAD in this case does estimate the privacy risk to be 0, whereas ReRo reports notably higher values (0.81 for Census and 0.73 for Texas). This underlines how RAD is the more reliable measure of actual privacy risks: RAD shows the absence of leakage when the attack's success relies solely on imputation, whereas ReRo suggests serious disclosures (or: attack potential), effectively overestimating the privacy risk.

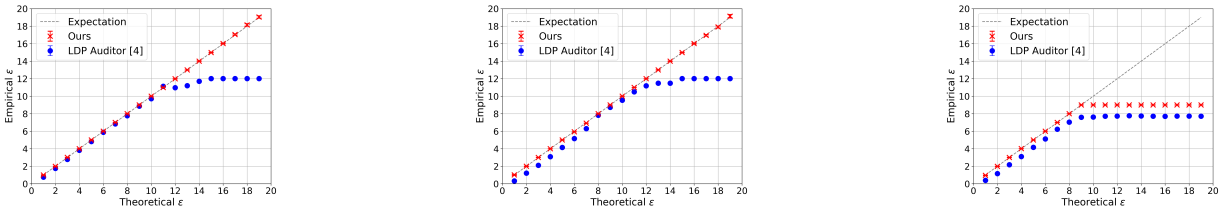
This tendency of ReRo to overestimate risk is not confined to this setting. In our optimal attacks on DP-SGD (Figure 3), ReRo consistently overestimates leakage across all investigated cases, with the effect becoming more pronounced as more auxiliary information is incorporated. Membership inference ($a(z) = z$) provides the clearest example, where ReRo reports risk values exceeding 0.6 even for privacy budgets $\varepsilon \leq 4$, which are commonly considered to offer strong privacy guarantees [46]. This behavior aligns with expectations, as ReRo cannot discount auxiliary information; consequently, greater attacker knowledge leads to larger overestimation.

Similarly, Figure 4 shows that ReRo fails to capture the effect of the success threshold η. As η increases, an oblivious attacker's success probability rises, but ReRo cannot account for this since it depends only on success probability and thus converges to 1 for all ε. This results in substantial overestimation: for $\eta = 100$, a trivial setting where any guess is correct, ReRo reports maximal risk despite the mechanism providing no advantage. In contrast, RAD properly discounts this effect, showing that increasing η boosts advantage only up to a point (here, $\eta = 40$), after which the advantage decreases as success becomes nearly granted.

7.3.3 Bound tightness. Figure 4 shows the results of RAD and ReRo for our optimal attack against Laplace mechanism on Adult including their corresponding theoretical bounds. Figure 6 shows the analogous for LDP mechanisms, GRR, OUE and SS, on the Porto dataset. On the x-axis, we see ε and the y-axes the exact estimated risk for such ε selection. Note that for LDP, RAD and ReRo results coincide, since the attack relies solely on the released output (with no auxiliary information or imputation effects). Moreover, the prior-based chance level under the uniform prior is negligible for $|\mathcal{Z}| = 3,052$. We therefore report only RAD to avoid redundancy.

We observe that our bounds (cf. Theorem 4.3) are tight for every prior π and capture even subtle differences between mechanisms. In particular, the RAD estimates for GRR perfectly match our perfect-reconstruction black-box bound (Theorem 5.3), confirming its tightness. Analogous results for the Geolife dataset are reported in the long version of this paper.

Moreover, Figure 4 clearly illustrates the impact of the data distribution: the skewed distribution (Figure 4c) constitutes the worst case, while the empirical distribution represents the best



(a) Gen. Randomized Response (GRR). (b) Subset Selection (SS). (c) Optimized Unary Encoding (OUE).

Figure 5: LDP Audit results from RAD-based auditing and LDP Auditor [6] on Porto dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

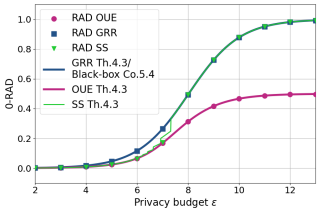


Figure 6: RAD results for LDP mechanisms (Porto). Lines show theoretical bounds and markers empirical RAD.

case. This highlights that knowledge of the data distribution can substantially improve utility; in the absence of such knowledge, the only safe choice is calibration with respect to the worst case.

Finally, these results provide concrete evidence for the importance of attack-based noise calibration. For identical values of ε, OUE offers significantly stronger protection against DRAs than GRR and SS. Hence, ε alone does not capture the full privacy picture, and RAD is essential for understanding the actual privacy implications of a mechanism for users.

7.3.4 Auditing Local DP with RAD. Figure 5 shows the results from our LDP Auditing experiments using the Porto dataset (experiments on the Geolife dataset yield similar results, which we show in the long version). They compare the accuracy of predicting the actual ε using our RAD-based auditing versus LDP Auditor. The closer the empirical ε is to the theoretical value (diagonal line), the more accurate the auditing tool. Additionally, smaller standard deviations indicate greater stability of the method.

For all tested mechanisms, our auditing approach improves over LDP Auditor for all ε values. In particular, we see that the highest ε LDP Auditor manages to estimate for both GRR and SS are capped around $\widetilde{\varepsilon} \approx 12.25$, hence preventing auditing of deployments with higher values. This limitation was already acknowledged by the authors of LDP Auditor, as it stems from the intrinsic shortcomings of the Clopper-Pearson method underlying their approach [6]. In contrast, the tightness of our RAD bound enables our auditing approach to accurately estimate empirical privacy budgets for the whole range, without such a limitation. Notably, for GRR and SS, our DP auditing yields near-perfect estimates for all epsilon values. For the OUE mechanism, our approach also outperforms LDP Auditor, however, the estimation accuracy declines at $\varepsilon \leq 9$. Note that this is an inherent limitation of OUE auditing as already mentioned in [6]: as we prove in Example 4.5, 0-RAD converges to $\frac{m-1}{2m}$ when ε tends to infinity. Overall, these results support that the universal tightness of our theoretical bound Theorem 4.3 enables precise and reliable auditing based on DRAs.

8 CONCLUSION

In this paper, we investigate the reconstruction risk that users incur when their data are processed by DP mechanisms. Our results reveal that the current state-of-the-art risk metric, ReRo [9], drastically overestimates the actual leakage of DP mechanisms when target-specific public knowledge exists—leading to excessive utility loss if used as noise calibration methods. Crucially, we show that under real attacks, existing ReRo bounds are violated.

To address these limitations, we first introduce η-RAD, a novel metric consistent with attribute and membership advantage, that accurately captures the privacy risk imposed by any specific mechanism. More importantly, we advance the understanding and practical interpretation of DP guarantees by proving tight bounds that connect DP mechanisms with their risk, using RAD. Offering new insights and clarity beyond existing analyses, we establish (i) universally tight bounds when the attacker's knowledge is specified, along with optimal strategies achieving them, (ii) closed-form bounds that remain valid regardless of auxiliary knowledge, and (iii) black-box upper bounds for settings with completely secret records. Our theoretical and empirical evaluation—across private learning, DP aggregation and LDP settings—demonstrates not only the robustness of RAD as a risk measure, but also the significant impact of our bounds on improving DP noise calibration (proving better utility) and auditing in DP (broadening the scope and improving accuracy).

Overall, our work demonstrates that privacy risk depends on the mechanism's structure, not just its nominal privacy parameters, and provides both fundamental insight and practical tools for privacy risk assessment and calibration—enabling notable utility gains without increasing the effective privacy risk.

ACKNOWLEDGMENTS

This work was funded by the Topic Engineering Secure Systems of the Helmholtz Association and supported by KASTEL SRL, and Germany's Excellence Strategy (EXC 2050/2 'CeTI'; ID 390696704). H.H.A. was supported by the French National Research Agency, under contracts: "ANR-24-CE23-6239" and "ANR-23-IACL-0006".

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. Association for Computing Machinery, New York, USA, 308–318. https://doi.org/10.1145/2976749.2978318

[2] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2867. https://doi.org/10.1145/3219819.3226070

[3] Meenatchi Sundaram Muthu Selva Annamalai, Borja Balle, Jamie Hayes, Georgios Kaissis, and Emiliano De Cristofaro. 2025. The Hitchhiker's Guide to Efficient, End-to-End, and Tight DP Auditing. arXiv:2506.16666

[4] Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. 2025. Nearly tight black-box auditing of differentially private machine learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vol. 37. Curran Associates Inc., Vancouver, BC, Canada, 131482–131502. https://doi.org/10.52202/079017-4179

[5] Héber H. Arcolezi. 2024. *LDP-Audit GitHub Repository*. GitHub. Retrieved March 9, 2026 from https://github.com/hharcolezi/ldp-audit

[6] Héber H. Arcolezi and Sébastien Gambs. 2024. Revealing the True Cost of Locally Differentially Private Protocols: An Auditing Perspective. *Proceedings on Privacy Enhancing Technologies* 2024 (2024), 123–141. https://doi.org/10.56553/popets-2024-0110

[7] Héber H. Arcolezi, Sébastien Gambs, Jean-François Couchot, and Catuscia Palamidessi. 2023. On the Risks of Collecting Multidimensional Data Under Local Differential Privacy. *Proceedings of the VLDB Endowment* 16, 5 (2023), 1126–1139. https://doi.org/10.14778/3579075.3579086

[8] François Baccelli, Bartlomiej Blaszczyszyn, and Mohamed Kadhem Karray. 2024. *Random Measures, Point Processes, and Stochastic Geometry*. Inria, France. https://inria.hal.science/hal-02460214v2

[9] Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing Training Data with Informed Adversaries. In *Symposium on Security and Privacy*. IEEE, San Francisco, USA, 1138–1156. https://doi.org/10.1109/SP46214.2022.9833677

[10] Borja Balle and Yu-Xiang Wang. 2018. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, Stockholm, Sweden, 394–403. https://proceedings.mlr.press/v80/balle18a.html

[11] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20

[12] Daniel Bernau, Günther Eibl, Philip W. Grassal, Hannah Keller, and Florian Kerschbaum. 2021. Quantifying identifiability to choose and audit ϵ in differentially private deep learning. *Proceedings of the VLDB Endowment* 14, 13 (2021), 3335–3347. https://doi.org/10.14778/3484224.348423

[13] Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, and Martin Vechev. 2021. DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers. In *Symposium on Security and Privacy (SP)*. IEEE, San Francisco, USA, 394–403. https://doi.org/10.1109/SP40001.2021.00081

[14] Yingyi Bu, Ada Wai Chee Fu, Raymond Chi Wing Wong, Lei Chen, and Jiuyong Li. 2008. Privacy preserving serial data publishing by role composition. *Proceedings of the VLDB Endowment* 1, 1 (2008), 845–856. https://doi.org/10.14778/1453856.1453948

[15] Mark Bun, Damien Desfontaines, Cynthia Dwork, Moni Naor, Kobbi Nissim, Aaron Roth, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. 2021. *Statistical inference is not a privacy violation*. DifferentialPrivacy.org. Retrieved March 9, 2026 from https://differentialprivacy.org/inference-is-not-a-privacy-violation/

[16] CJ Carey, Travis Dick, Alessandro Epasto, Adel Javanmard, Josh Karlin, Shankar Kumar, Andres Muñoz Medina, Vahab Mirrokni, Gabriel Henrique Nunes, Sergei Vassilvitskii, and Peilin Zhong. 2023. Measuring re-identification risk. *Proceedings of the ACM on Management of Data* 1, 2 (2023), 1–26. https://doi.org/10.1145/3589294

[17] Konstantinos Chatzikokolakis, Giovanni Cherubin, Catuscia Palamidessi, and Carmela Troncoso. 2023. Bayes security: A not so average metric. In *36th Computer Security Foundations Symposium (CSF)*. IEEE Computer Society, Los Alamitos, CA, USA, 388–406. https://doi.org/10.1109/CSF57540.2023.00011

[18] Graham Cormode, Shripad Gade, Samuel Maddock, and Enayat Ullah. 2025. Synthetic Tabular Data: Methods, Attacks and Defenses. *Proceedings of the VLDB Endowment* 18, 12 (2025), 5448–5450. https://doi.org/10.14778/3750601.3750692

[19] Teddy Cunningham, Graham Cormode, Hakan Ferhatosmanoglu, and Divesh Srivastava. 2021. Real-world trajectory sharing with local differential privacy. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2283–2295. https://doi.org/10.14778/3476249.3476280

[20] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific Reports* 3 (2013), 1376. https://doi.org/10.1038/srep01376

[21] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. 2018. Detecting Violations of Differential Privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS'18)*. Association for Computing Machinery, New York, NY, USA, 475–489. https://doi.org/10.1145/3243734.3243818

[22] Jinshuo Dong, Aaron Roth, and Weijie J. Su. 2022. Gaussian Differential Privacy. *Journal of the Royal Statistical Society* 84, 1 (2022), 3–37. https://doi.org/10.1111/rssb.12454

[23] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*. Springer, Berlin, Heidelberg, 1–12. https://doi.org/10.1007/11787006_1

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference (Lecture Notes in Computer Science)*, Vol. 3876. Springer, Berlin, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[25] Cynthia Dwork and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, Vol. 9. Now Publishers, Hanover, MA, USA. https://doi.org/10.1561/0400000042

[26] Úlfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. 2019. That which we call private. arXiv:1908.03566

[27] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. Association for Computing Machinery, New York, USA, 1054–1067. https://doi.org/10.1145/2660267.2660348

[28] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15)*. Association for Computing Machinery, New York, USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[29] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *23rd USENIX Security Symposium*. USENIX Association, San Diego, CA, 17–32. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew

[30] Elena Ghazi and Ibrahim Issa. 2024. Total variation meets differential privacy. *IEEE Journal on Selected Areas in Information Theory* 5 (2024), 207–220. https://doi.org/10.1109/JSAIT.2024.3384083

[31] Daniele Gorla, Louis Jalouzot, Federica Granese, Catuscia Palamidessi, and Pablo Piantanida. 2025. On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting. *IEEE Transactions on Dependable and Secure Computing* 22, 5 (2025), 5494–5507. https://doi.org/10.1109/TDSC.2025.3568160

[32] Patricia Guerra-Balboa, Annika Sauer, and Thorsten Strufe. 2024. Analysis and Measurement of Attack Resilience of Differential Privacy. In *Proceedings of the 23rd Workshop on Privacy in the Electronic Society (WPES'24)*. Association for Computing Machinery, New York, USA, 155–171. https://doi.org/10.1145/3689943.3695046

[33] M. Emre Gursoy, Ling Liu, Ka-Ho Chow, Stacey Truex, and Wenqi Wei. 2022. An Adversarial Approach to Protocol Analysis and Selection in Local Differential Privacy. *IEEE Transactions on Information Forensics and Security* 17 (2022), 1785–1799. https://doi.org/10.1109/TIFS.2022.3170242

[34] Jamie Hayes, Borja Balle, and Saeed Mahloujifar. 2023. Bounding training data reconstruction in DP-SGD. In *Advances in Neural Information Processing Systems (NeurIPS'23)*, Vol. 36. Curran Associates, Inc., New Orleans, USA, 78696–78722. https://proceedings.neurips.cc/paper_files/paper/2023/file/f8928b073ccbec15d35f2a9d39430bfd-Paper-Conference.pdf

[35] Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. 2022. TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data. arXiv:2211.06550

[36] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2023. Investigating Membership Inference Attacks under Data Dependencies. In *IEEE 36th Computer Security Foundations Symposium (CSF'23)*. IEEE Computer Society, Los Alamitos, CA, USA, 473–488. https://doi.org/10.1109/CSF57540.2023.00013

[37] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private sgd?. In *Advances in Neural Information Processing Systems (CSF'23)*, Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 22205–22216. https://proceedings.neurips.cc/paper_files/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf

[38] Bargav Jayaraman. 2022. *Analyzing the Leaky Cauldron: Inference Attacks on Machine Learning*. Ph.D. dissertation. University of Virginia. https://doi.org/10.18130/myhy-tv46

[39] Bargav Jayaraman. 2022. *EvaluatingDPML GitHub Repository*. GitHub. Retrieved March 9, 2026 from https://github.com/bargavj/EvaluatingDPML

[40] Bargav Jayaraman and David Evans. 2022. Are Attribute Inference Attacks Just Imputation?. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*. Association for Computing Machinery, New York, NY, USA, 1569–1582. https://doi.org/10.1145/3548606.3560663

[41] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference*

on International Conference on Machine Learning (ICML'16). JMLR.org, New York, USA, 2436–2444. http://proceedings.mlr.press/v48/kairouz16.pdf

[42] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The Composition Theorem for Differential Privacy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML'15)*, Vol. 37. JMLR.org, New York, USA, 2436–2444. https://proceedings.mlr.press/v37/kairouz15.html

[43] Daniel Kifer, John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. 2022. Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census. arXiv:209.03310

[44] Bogdan Kulynych, Juan F Gomez, Georgios Kaissis, Flavio du Pin Calmon, and Carmela Troncoso. 2024. Attack-aware noise calibration for differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS'24)*, Vol. 37. Curran Associates, Inc., New York, USA, 134868–134901. https://doi.org/10.52202/079017-4286

[45] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[46] Jaewoo Lee and Chris Clifton. 2011. How Much Is Enough? Choosing ϵ for Differential Privacy. In *Information Security*. Springer, Berlin, Heidelberg, 325–340. https://doi.org/10.1007/978-3-642-24861-0_22

[47] Szilvia Lestyán, Gergely Ács, and Gergely Biczók. 2022. In Search of Lost Utility: Private Location Data. arXiv:2008.01665

[48] Yun Lu, Malik Magdon-Ismail, Yu Wei, and Vassilis Zikas. 2024. Eureka: A General Framework for Black-box Differential Privacy Estimators. In *Symposium on Security and Privacy (SP)*. IEEE, San Francisco, USA, 913–931. https://doi.org/10.1109/SP54263.2024.00166

[49] David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press, New York, NY, USA.

[50] Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. 2024. Auditing f-Differential Privacy in One Run. arXiv:2410.22235

[51] Mani Malek, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramèr. 2021. Antipodes of label differential privacy: PATE and ALIBI. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS '21)*, Vol. 34. Curran Associates Inc., Red Hook, NY, USA, 6934–6945. https://proceedings.neurips.cc/paper_files/paper/2021/file/37ecd27608480aa3569a511a638ca74f-Paper.pdf

[52] Frank D. McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '09)*. Association for Computing Machinery, New York, NY, USA, 19–30. https://doi.org/10.1145/1559845.1559850

[53] Sebastian Meiser. 2018. Approximate and Probabilistic Differential Privacy Definitions. Cryptology ePrint Archive, Paper 2018/277. https://eprint.iacr.org/2018/277

[54] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2023. What are the chances? explaining the epsilon parameter in differential privacy. In *Proceedings of the 32nd USENIX Conference on Security Symposium (SEC '23)*. USENIX Association, USA, 1613–1630. https://www.usenix.org/system/files/usenixsecurity23-nanayakkara.pdf

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

[56] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *Symposium on Security and Privacy (SP)*. IEEE, Los Alamitos, CA, USA, 866–882. https://doi.org/10.1109/SP40001.2021.00069

[57] Meghan O'Connell, Matias Moreira, and Wendy Kan. 2015. *ECML/PKDD 15: Taxi Trajectory Prediction (I)*. Kaggle. Retrieved March 9, 2026 from https://kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i

[58] OpenStreetMap contributors. 2017. *Planet dump retrieved from https://planet.osm.org*. OSM. Retrieved March 9, 2026 from https://www.openstreetmap.org

[59] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. What Does The Crowd Say About You? Evaluating Aggregation-based Location Privacy. *Proceedings on Privacy Enhancing Technologies* 2017 (2017), 156–176. https://doi.org/10.1515/popets-2017-0043

[60] Timothy Sauer. 2011. *Numerical Analysis*. Addison-Wesley Publishing Company, USA.

[61] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez. 2014. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal* 23, 5 (2014), 771–794. https://doi.org/10.1007/s00778-014-0351-4

[62] Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2023. Privacy auditing with one (1) training run. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS '23)*, Vol. 36. Curran Associates Inc., Red Hook, NY, USA, 49268–49280. https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6f6e0d6781d1cb8689192408946d73-Paper-Conference.pd

[63] Latanya Sweeney. 2000. *Simple Demographics Often Identify People Uniquely*. Data Privacy Working Paper 3. Carnegie Mellon University, Data Privacy Lab. https://dataprivacylab.org/projects/identifiability/paper1.pdf

[64] TensorFlow contributors. 2025. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved March 9, 2026 from https://www.tensorflow.org

[65] Florian Tramèr, A. Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. 2022. Debugging Differential Privacy: A Case Study for Privacy Auditing. arXiv:2202.12219

[66] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium*. USENIX Association, Vancouver, BC, 729–745. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao

[67] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747

[68] Yonghui Xiao and Li Xiong. 2015. Protecting Locations with Differential Privacy under Temporal Correlations. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1298–1309. https://doi.org/10.1145/2810103.2813640

[69] Min Ye and Alexander Barg. 2018. Optimal Schemes for Discrete Distribution Estimation Under Locally Differential Privacy. *IEEE Transactions on Information Theory* 64, 8 (2018), 759–763. https://doi.org/10.1109/ISIT.2017.8006630

[70] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *31st Computer Security Foundations Symposium (CSF'18)*. IEEE, Los Alamitos, CA, USA, 268–282. https://doi.org/10.1109/CSF.2018.00027

[71] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. *Geolife GPS trajectory dataset - User Guide*. Microsoft. Retrieved March 9, 2026 from https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/