Karlsruhe Institute of Technology

PRIVACY AND SECURITY

Measurement of Attack Resilience of Differential Privacy

Master's Thesis of

Annika Sauer

at the Department of Informatics
KASTEL – Institute of Information Security and Dependability
Practical IT Security

Reviewer:	Prof. Dr. Thorsten Strufe
Second reviewer:	Prof. Dr. Jörn Müller-Quade
Advisor:	M. Sc. Patricia Guerra-Balboa
Second advisor:	Dr. Héber Hwang Arcolezi

10 March 2025 – 10 September 2025

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Abstract

Differential Privacy (DP) is a leading framework for privacy-preserving data analysis, offering formal guarantees controlled by the privacy budget parameter ε. In practice, however, it is up to practitioners to select ε—often without clear guidance on how this choice impacts protection against privacy attacks such as attribute inference and data reconstruction. There is a need for metrics that capture the privacy risks posed by concrete adversaries to support such decisions.

Existing metrics like Reconstruction Robustness (ReRo) aim to quantify this risk but potentially overestimate information leakage, leading to overly conservative noise calibration and unnecessary utility loss.

A new metric, Unbiased Reconstruction Robustness (U-ReRo), was proposed to address this issue. It is a refined metric that aims to distinguish true privacy leakage from adversarial prior knowledge and data correlation. While U-ReRo provides tighter bounds than ReRo in theory, its practical performance has not yet been evaluated. It is therefore unclear whether U-ReRo improves over ReRo and whether the provided bounds are tight in practice.

In this work, we conduct empirical evaluations using realistic attacks on DP mechanisms to compare ReRo and U-ReRo. Our results show that ReRo overestimates privacy risk, especially when attacks exploit background knowledge or imputation, whereas U-ReRo more accurately reflects actual leakage. We further prove that our bound on U-ReRo under pure DP and uniform prior is tight in settings without target-specific auxiliary knowledge, providing practical guidance for DP parameter selection.

Beyond risk estimation, we apply U-ReRo to DP auditing—the task of empirically estimating the privacy loss ε from the outputs of a DP mechanism. This task is critical for validating the correctness of deployed DP mechanisms, as theoretical guarantees may not capture the true behavior of practical implementations. However, prior auditing methods have focused predominantly on membership inference attacks. In contrast, U-ReRo extends auditing beyond this setting, offering a framework for estimating privacy loss more broadly. In the Local DP setting, U-ReRo furthermore achieves more accurate auditing results than the state-of-the-art.

Zusammenfassung

Differential Privacy (DP) ist das führende Konzept für die datenschutzfreundliche Datenanalyse und bietet formale Garantien, die durch den Parameter ε kontrolliert werden. In der Praxis ist es jedoch Praktikern überlassen, ε auszuwählen - oft ohne klare Anweisungen, wie sich diese Wahl auf den Schutz vor Angriffen auf die Privatsphäre auswirkt, wie z. B. Attributinferenz oder Datenrekonstruktion. Um diese Risiken zu erfassen und die Entscheidungen von Praktikern zu unterstützen, benötigen wir Metriken, die die von konkreten Angreifern ausgehenden Risiken für die Privatsphäre erfassen.

Bestehende Metriken wie Reconstruction Robustness (ReRo) zielen darauf ab, dieses Risiko zu quantifizieren, überschätzen aber möglicherweise Informationsleaks, was Praktiker dazu verleiten kann, ε zu konservativ zu wählen, was zu unnötigem Nutzenverlust führt, ohne dass Privatsphäre dadurch besser geschützt wird.

Dieses Problem wird von einer neuen Metrik, Unbiased Reconstruction Robustness (U-ReRo), addressiert. U-ReRo ist eine verfeinerte Metrik, die darauf abzielt, echte Informationsleaks von Wissen zu unterscheiden, über das ein Angreifer bereits verfügte oder, das aus Datenkorrelation ableitbar ist. Während U-ReRo in der Theorie Informationsleaks präziser misst als ReRo, wurde seine praktische Leistung noch nicht bewertet.

In dieser Arbeit vergleichen wir ReRo und U-ReRo durch empirische Evaluierung von Angriffen auf DP-Mechanismen. Unsere Ergebnisse zeigen, dass ReRo das Risiko für die Privatsphäre überschätzt, insbesondere wenn Angriffe Hintergrundwissen oder Imputation ausnutzen, während U-ReRo das tatsächliche Informationsleak genauer wiedergibt. Wir beweisen außerdem, dass unsere theoretische Schranke für U-ReRo präzise ist, wenn Angreifer über keinerlei Hintergrundwissen verfügen, was eine praktische Anleitung für die Auswahl der DP-Parameter bietet.

Über die Risikoabschätzung hinaus wenden wir U-ReRo auf das DP-Auditing an – die Aufgabe, den Verlust an Privatsphäre aus den Ausgaben eines DP-Mechanismus empirisch zu schätzen. Diese Aufgabe ist entscheidend für die Validierung der Korrektheit von eingesetzten DP-Mechanismen, da theoretische Garantien möglicherweise nicht das tatsächliche Verhalten praktischer Implementierungen erfassen. Bisherige Prüfmethoden haben sich jedoch hauptsächlich auf Angriffe durch Mitgliedschaftsinferenz konzentriert. Im Gegensatz dazu erweitert U-ReRo das Auditing über diesen Rahmen hinaus und bietet einen Rahmen für die Abschätzung des Verlusts an Privatsphäre im Kontext weiterer Angriffe. Im Kontext von Lokaler DP erzielt U-ReRo außerdem genauere Auditing-Ergebnisse als frühere Ansätze.

Contents

List of Figures

List of Tables

1. Introduction

Data analysis plays a critical role in domains such as medical research and urban planning, where insights are often drawn from datasets containing sensitive personal information such as mobility traces or health records [80]. Even when datasets are anonymized or shared in aggregate form, real-world attacks have shown that individual information can still be inferred, showcasing that removing direct identifiers does not ensure privacy [86, 71, 68]. These findings have motivated the development of privacy-preserving data analysis techniques designed to limit the risk of such attacks. Among these, differential privacy (DP) [27] has emerged as the de facto standard.

DP is a robust framework that allows us to learn statistics about the population while providing measurable privacy leakage. The privacy leakage of DP is parameterized by the *privacy budget* ε [27]. It controls how similar the probabilities of observing the same output are, independent of whether an individual participated in the data collection or not. The choice of this parameter is key. If ε is too large, private information will be disclosed. Choosing ε too small can significantly reduce the usefulness of the mechanism's output for data analysis [29]. Since ε ranges from 0 to infinity, understanding its practical implications is difficult: while it is known that smaller values offer stronger privacy, this provides little guidance for practitioners trying to balance utility and protection.

One promising direction to address this gap is to evaluate how the selection of ε affects a system's vulnerability to privacy attacks. That is, we can select the privacy budget according to its effectiveness in limiting what an adversary can infer about individuals in the dataset. In this context, recent research has analyzed the success of inference attacks under different privacy budgets, e.g. [84, 97, 31, 12, 10], leading to theoretical bounds on the adversary's capabilities. This approach allows to calibrate ε based on the types of privacy risk that have to be mitigated.

While DP provides strong theoretical guarantees [29, 28], real-world implementations can deviate due to misparameterization or implementation errors [45, 18]. *DP auditing* complements formal analysis by empirically testing privacy—often via simulated or real attacks–to measure information leakage [45, 5]. This is increasingly important for deployed systems, such as the 2020 U.S. Census [3], especially when full system access is unavailable.

Implementing DP then requires a calibrated choice of ε which in turn depends on metrics to quantify adversarial success. In this context, the metric of *Reconstruction Robustness* (ReRo), introduced in [10], allows to quantify the adversary's success in recovering information about individuals from DP mechanisms for different inference attacks. ReRo measures the probability that an adversary can correctly reconstruct their target record from the output of a DP mechanism.

Importantly, DP is known to upper-bound ReRo, meaning that for any mechanism satisfying DP, there exists a theoretical limit on how well any adversary can perform reconstruction [10]. This link between DP and ReRo allows practitioners to relate the

abstract privacy budget ε to a concrete adversarial success metric, enabling more informed decisions when choosing ε.

However, previous work [37] discusses theoretical limitations of the ReRo metric. While ReRo is proposed as a general-purpose measure for reconstruction risk, it diverges from earlier attack metrics when applied to specific inference settings. Most notably, existing metrics for membership and attribute inference operate as *advantages*: they measure the difference between an adversary's success on individuals who participated in the dataset and those drawn from the underlying data distribution [97]. This difference is crucial, as it isolates the privacy leakage attributable to data participation, filtering out success that stems from other sources. ReRo, by contrast, reports the raw success probability of the adversary, without any baseline correction. This leads to a key limitation: it can overestimate privacy leakage in scenarios where the adversary's performance relies on background knowledge or imputation. For example, an attacker might accurately infer that a patient has lung cancer based on the fact that they are a smoker, independent of whether their medical record was included in the dataset [17]. This is a classic instance of the *privacy fallacy* [17], where the inference appears to violate privacy but in fact exploits public information.

To address the bias inherent in ReRo, Guerra-Balboa et al. introduced a refined metric: *Unbiased Reconstruction Robustness* (U-ReRo) [37]. U-ReRo extends ReRo by subtracting a correction term that accounts for adversarial success on records drawn from the underlying distribution. Thus far, however, U-ReRo has only been motivated theoretically, and no empirical evaluation is available. Without such validation, it remains unclear whether U-ReRo effectively corrects the overestimation problem of ReRo or whether its bounds are sufficiently tight.

A more fundamental limitation is that U-ReRo–like its predecessor ReRo–does not consider *target-specific* auxiliary knowledge *aux*. Ignoring this knowledge means that both U-ReRo and ReRo only model adversaries with either no background knowledge or general, population-level knowledge. Many important attacks lie outside this restricted threat model: in *attribute inference*, the adversary knows some public attributes of a target record and attempts to infer a sensitive attribute; in *membership inference*, the adversary knows the entire target record and tries to determine whether it is included in a given dataset. Neither setting can be accurately modeled by ReRo or U-ReRo, as both metrics assume the adversary has no access to *aux*.

Overall, while U-ReRo is a promising metric to guide privacy budget selection, it remains unverified whether it reliably captures actual privacy leakage or whether its bounds offer meaningful worst-case guarantees. In addition, the metric must be extended to account for *aux* in order to capture broader and more realistic threat scenarios.

The first contribution of this thesis is therefore an empirical evaluation of U-ReRo and a comparison with ReRo. We show that in privacy attacks across different domains, such as Local Differential Privacy (LDP) [25, 51] queries and differentially private machine learning (ML) [1], ReRo overestimates the privacy leakage whereas U-ReRo detects when adversarial success is due to background knowledge or imputation. Moreover, in genuine leakage scenarios, both metrics yield similar results—demonstrating that U-ReRo remains sensitive to actual privacy risks, producing high values when leakage is present and low values when it is not.

Guerra-Balboa et al. furthermore proved that DP implies an upper bound on U-ReRo [37]. While this theoretical bound provides a useful worst-case guarantee, it remains unclear whether it is tight in practice. Understanding the tightness of this bound is crucial: if the bound is overly loose, it may significantly overestimate the risk of reconstruction. An overestimation of the risk, in turn, leads to overly low choices of the privacy budget, resulting in a loss of utility without a corresponding gain in privacy.

The second contribution of this thesis then is the empirical evaluation of theoretical upper bounds on U-ReRo. We improve upon the theoretical bound presented in [37] in settings where the adversary has no target-specific auxiliary knowledge. We furthermore show that our refined bounds are perfectly tight under uniform prior – that is, the empirical performance of reconstruction attacks matches the bound exactly. Even under non-uniform data distributions, our novel bounds provide a considerable improvement over the state-of-the-art and tightly capture the privacy leakage for low values of the privacy budget.

As a third contribution, we extend U-ReRo to scenarios where the adversary has access to *aux*, introducing a novel metric: *Aux-Aware U-ReRo*. This metric provides a truly general framework for data reconstruction attacks. By varying the type of aux available to the adversary—ranging from the entire record (membership inference), to selected public attributes (attribute inference), or no auxiliary information (data reconstruction)—Aux-Aware U-ReRo captures a broader spectrum of realistic threat scenarios. We furthermore provide a first bound for Aux-Aware U-ReRo under DP and show it to be reasonably tight for low- to mid-range privacy budgets.

Beyond understanding the theoretical and empirical behavior of U-ReRo, we also explore its potential as a tool for *DP auditing*. As with the broader study of attacks on DP, the DP auditing literature has primarily focused on auditing via membership inference attacks, e.g. [23, 14, 72, 85]. While some recent works have explored auditing through attribute inference and data reconstruction attacks [43, 63, 61], this remains a comparatively underexplored area.

However, membership is not always sensitive–for example, in social media platforms, it is typically public [10]. In such contexts, membership inference attacks may not pose a meaningful threat, while data reconstruction or attribute inference could still represent significant privacy risks. In these scenarios, it is of interest to enforce protection specifically against reconstruction attacks, which might allow for selecting higher privacy budgets and thus improving utility [10]. To support this, auditing tools should allow to address these broader threat models as well.

Building on our result that the U-ReRo bound is perfectly tight under a uniform prior–an assumption made in previous works on DP auditing [6, 13]–we propose U-ReRo as an auditing metric that extends auditing beyond membership inference to data reconstruction. As the third contribution of this thesis, we introduce a basic framework for U-ReRo-based auditing and demonstrate that it achieves performance comparable to, and in some cases exceeding, the state-of-the-art in the local DP setting.

Our main contributions are:

- We demonstrate that ReRo does not account for the privacy fallacy. Attack success that results from the adversary's background knowledge or from learnt global statistics is conflated with true privacy leakage resulting from individual participation in

the data collection. In consequence, ReRo overestimates the privacy leakage and can thus lead to an overly conservative choice of the privacy budget.

- We show that U-ReRo distinguishes actual privacy leakage from attack success based on the adversary's background knowledge or imputation. U-ReRo then provides a tighter estimate of the true risk a user is exposed to by sharing their data and allows to calibrate the privacy budget accordingly.
- We show that our theoretical bound on U-ReRo is perfectly tight when the adversary has no background knowledge. In this case, we can present an attack that achieves the predicted worst-case performance.
- We introduce a novel metric, Aux-Aware U-ReRo, that extends U-ReRo to settings where the adversary has access to target-specific auxiliary knowledge. To the best of our knowledge, we are the first to provide a metric for data reconstruction that is truly general and allows to model both membership and attribute inference as special cases. We also provide a first bound on Aux-Aware U-ReRo and show that it is near tight for low privacy budgets.
- We show that U-ReRo can be used as an effective tool for DP auditing in the LDP setting where it provides tighter estimates of the privacy budget compared to the state-of-the-art.

2. Related Work

To address the challenge of calibrating ε, researchers have investigated how well DP mechanisms withstand concrete privacy attacks. A common approach is to assess or bound the success of such attacks–either empirically or theoretically–and relate it to the privacy budget ε. These efforts aim to clarify the real-world protections that specific values of ε afford by quantifying the extent to which adversaries can succeed under DP guarantees.

2.1. DP Bounds on Privacy Attacks

The three dominant types of attacks studied under DP are membership inference, attribute inference, and data reconstruction [37]. Among these, membership inference has received the most attention. The goal of a membership inference attack is to determine whether a particular record was part of a private dataset [97]. Yeom et al. were the first to establish a theoretical bound on the effectiveness of membership inference under DP [97]. They introduce the notion of the *membership advantage*–defined as the difference between the true positive rate and the false positive rate of an attack–as a metric to quantify an adversary's success in distinguishing between records that were part of the training dataset and those that were not. For any ε-DP mechanism, they prove that the membership advantage is upper-bounded by DP. This bound was later extended in [31] to the more general (ε, δ)-DP setting. Most recently, Humphries et al. proposed a tighter bound on membership advantage that further improves over previous results [44]. Closely related are attribute inference attacks, in which the adversary has partial information about a target record and attempts to infer unknown sensitive attributes [97]. Yeom et al. showed that attribute inference can be reduced to membership inference and proposed the *attribute advantage* as an analogous metric, proving that it too is bounded by DP. This perspective–that attribute inference is constrained by membership inference–has since been supported by further work [98].

However, these bounds are explicitly formulated for the membership and attribute advantages and are thus specific to membership and attribute inference attacks. They are not directly applicable to broader attack settings.

Data reconstruction attacks represent such a broader setting: rather than inferring sensitive attributes or membership status, the adversary seeks to reconstruct entire records [10]. This formulation allows to encompass both membership and attribute inference as special cases–an adversary with partial knowledge of a record's attributes, for instance, reduces the reconstruction task to attribute inference [10].

To quantify the effectiveness of data reconstruction attacks, Balle et al. introduce the concept of *Reconstruction Robustness* (ReRo), which measures the probability that an

adversary can produce an accurate reconstruction of a target record. They further provide a theoretical upper bound on ReRo under DP [10]. Building on this foundation, [37] extends these insights to derive general theoretical bounds for attribute inference attacks. Complementarily, Guo et al. establish a tighter bound on the mean squared error of adversarial estimates of training samples under DP [38]. Furthermore, the theoretical understanding of data reconstruction is refined for the private learning algorithm DP-SGD [1] in [42], where the authors demonstrate their bound to be nearly tight for this specific setting. More recently, Cummings et al. note that most reconstruction analyses assume an adversary with full knowledge of the training database except one record, and propose a relaxation, Distributional ReRo, where the adversary only has access to population-level information.

Prior approaches grounded in the ReRo framework establish a theoretical foundation for analyzing data reconstruction attacks; however, they exhibit a critical limitation: ReRo quantifies the adversary's reconstruction success probability without discerning whether this success arises from actual privacy leakage. It is essential to differentiate between individual privacy leakage and statistical imputation – learning population-level statistics. Specifically, we do not consider learning from global distributional properties as a privacy violation [64].

Guerra-Balboa et al. propose U-ReRo as a solution to this overestimation issue: it builds directly on ReRo and subtracts a correction term to account for adversarial success based on background knowledge or imputation. To date, it remains purely theoretical, with no empirical evidence to demonstrate its effectiveness or confirm that the proposed bounds are tight in practice [37].

While U-ReRo is a promising direction as it addresses the potential overestimation of privacy leakage of ReRo, it inherits another fundamental limitation of ReRo: While data reconstruction attacks explicitly account for target-specific auxiliary knowledge *aux*, ReRo does not [10]. Despite its claimed generality, this omission means that ReRo, and all metrics derived from it, cannot accurately measure attack success in many important settings, such as attribute inference or membership inference, where *aux* is essential.

In summary, prior work on quantifying the success of privacy attacks under DP has primarily focused on specific attack classes, such as membership or attribute inference. While these studies provide valuable insights, their applicability is limited to these specific attacks or to particular settings–such as assumptions about data dependencies [89] or specific algorithms like DP-SGD [42]. Among the existing theoretical approaches, bounds on data reconstruction have emerged as the most general and promising, as many attacks–including membership and attribute inference–can be framed as reconstruction problems [37]. However, the dominant metric for evaluating reconstruction risk, ReRo [10], both potentially overestimates privacy leakage and ignores target-specific auxiliary knowledge, meaning it may provide an overly pessimistic assessment of attack success and cannot capture attacks such as attribute inference or membership inference, despite its claim of generality.

2.2. DP Auditing

While the study of privacy attacks under DP focuses on providing theoretical guarantees, DP auditing as introduced by [45] takes a complementary approach: it seeks to measure or estimate the effective privacy guarantees of a system, often in a black-box or post-deployment setting. Auditing methods assess whether the privacy loss incurred by a mechanism aligns with its claimed privacy budget [45].

To date, most DP auditing approaches have focused on membership inference attacks. These methods evaluate whether individual data points–often specially crafted canaries–can be detected as part of the training data. Notable examples include auditing through prediction loss [5], shadow models [69, 70, 20], and gradient-based methods in federated learning [60, 61, 85]. Auditing has been applied most extensively to DP-SGD [1], with studies demonstrating both tight estimates of the privacy budget [72] and discovering implementation flaws [88].

However, focusing solely on membership inference presents a limitation: an individual's membership in a dataset is not always the sensitive information. In many real-world cases, the fact that an individual participated in the data collection is public, while specific attributes or contents are sensitive [10]. Thus, to fully assess privacy risk, auditing methods should also account for attacks beyond membership inference.

Only a few recent efforts move in this direction: for attribute inference, [63] propose an auditing setup for label-DP while [43] incorporate attribute inference into a broader synthetic data auditing framework. For data reconstruction, [61] adapt the setup from [42] to reconstruct training examples from a candidate pool. However, these approaches remain narrow in scope—attribute inference strategies target highly specific setups (e.g., label-DP or synthetic data), and reconstruction-based auditing neither reconstructs from scratch nor generalizes beyond classification tasks. As such, existing auditing methods do not provide a comprehensive view of privacy leakage beyond membership inference.

3. Background

This chapter presents the core concepts for this work. We start by formalizing our understanding of DP and data reconstruction. Then, we present the concepts of ReRo and U-ReRo as performance metrics for data reconstruction attacks. An overview of our formal notations can be found in Table 3.1.

3.1. Differential Privacy

DP was first introduced as a privacy notion for statistical databases. Intuitively, it states that the participation of a user in a database has only a small impact on their privacy [29]. Hence, the ability of an adversary to infer private information about the user is limited. This limit on inference ability is established as follows:

Definition 1 ((ε, δ)-Differential Privacy [29]) *A randomized mechanism $\mathcal{M} : \mathcal{Z}^n \to \Theta$ is differentially private if for all $S \subseteq \Theta$ and for every pair of databases $D, D' \in \mathcal{Z}^n$ such that $d_H(D, D') \leq 1$:*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

where d_H denotes the Hamming distance [41].

For two *neighboring* databases D, D' (databases that differ exactly in one entry, i.e. $d_H(D, D') \leq 1$), DP states that the probability of observing a certain outcome has to be similar. The degree of similarity is parametrized by the privacy budget ε. It controls how close the probabilities of observing the same output on databases D and D' must be. A smaller ε corresponds to a stronger guarantee. The DP parameter δ captures the idea that the postulated guarantee may not hold in all cases [27]. When $\delta = 0$, we speak of *pure* DP, else of *approximate* DP.

The information that DP protects depends on the definition of neighborhood of databases D, D'. We consider *bounded* DP where D' is obtained from D by replacing one user's complete record with that of another [67].

The classic definition of DP as presented above assumes the existence of a trusted data collector that receives the data of all users, applies some mechanism $\mathcal{M}$ to the collected data and then publishes the obfuscated data [27]. However, the existence of such a trusted collector cannot be assumed for all settings. Instead, the task of sanitizing the data can also be distributed to each user before sending their data. The original data is then stored only by the user who sends their sanitized data to be collected [28]. This is called LDP [25], formally defined in Definition 2.

Notation	Meaning
$\mathcal{Z}$	Data domain
$m = \lvert\mathcal{Z}\rvert$	Size of the data domain
$z \in \mathcal{Z}$	Data point
π	Data distribution
$\mathcal{Z} \sim \pi$	Random variable following distribution of data records π
$\mathcal{Z}^n$	Space of datasets of size n
D	Dataset
D_-	Dataset missing one record
$\mathcal{M}\colon \mathcal{Z}^n \to \Theta$	DP mechanism
$\mathcal{M}(D)$	Dataset perturbed with mechanism $\mathcal{M}$
$\theta \sim \mathcal{M}(D)$	Random variable following the distribution of output of the mechanism on D
A	Attack
aux	Target-specific auxiliary knowledge on target $z \in \mathcal{Z}$

Table 3.1.: Table of notation.

Definition 2 (ε-Local Differential Privacy [25]) *A randomized mechanism $\mathcal{M} : \mathcal{Z} \to \Theta$ satisfies ε-local differential privacy if for all $S \subseteq \Theta$ and for all pairs of input records $z_0, z_1 \in \mathcal{Z}$:*

$$\Pr[\mathcal{M}(z_0) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(z_1) \in S].$$

Intuitively, LDP ensures that the output distribution of $\mathcal{M}$ is nearly indistinguishable regardless of the user's true input.

DP and LDP focus on bounding the ratio of probabilities of outputs for neighboring inputs, which is equivalent to controlling the type I and type II errors of a hypothesis test distinguishing those inputs [24]. Recently, a more general framework called *f-differential privacy* (f-DP) has been introduced [24]. Instead of specifying a single parameter ε, f-DP characterizes the privacy guarantee via a *trade-off function* $f : [0,1] \to [0,1]$ that describes the minimal achievable type II error β as a function of the type I error α for any hypothesis test distinguishing two neighboring datasets.

Definition 3 (Trade-off function [24]) *For any two probability distributions P and Q, the trade-off function $T(P,Q) : [0,1] \to [0,1]$ is defined as*

$$T(P,Q)(\alpha) := \inf_{\phi} \left\{ \Pr_{Q}[\phi(X) = 0] \;\middle|\; \Pr_{P}[\phi(X) = 1] \leq \alpha \right\},$$

where ϕ is a hypothesis test for testing $H_0 : X \sim P$ versus $H_1 : X \sim Q$, and $\alpha \in [0,1]$ specifies the type I error of the test.

$T(P,Q)(\alpha)$ describes the minimal type II error (false negative) achievable by any test distinguishing P from Q, given that the type I error (false positive) is bounded by α. Based on the trade-off function, f-DP is defined as follows:

Definition 4 (f-Differential Privacy [24]) *A randomized mechanism $\mathcal{M}$ satisfies f-DP if, for all neighboring datasets D, D', the trade-off function between the distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is lower-bounded by f:*

$$T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) \geq f(\alpha), \quad \forall \alpha \in [0, 1].$$

Intuitively, f-DP provides a complete characterization of the trade-off between false positives and false negatives for distinguishing neighboring inputs. It generalizes traditional (ε, δ)-DP; every (ε, δ)-DP mechanism satisfies f-DP with trade-off function [24]

$$f_{\varepsilon,\delta}(\alpha) = \max\{0,\ 1 - \delta - e^{\varepsilon}\alpha,\ e^{-\varepsilon}(1 - \delta - \alpha)\}, \quad \alpha \in [0, 1].$$

The f-DP representation enables the computation of the *total variation* (TV) which quantifies the maximum difference in output distributions of a mechanism on neighboring datasets. It thus provides a measure of distinguishability and privacy loss.

Definition 5 (Total Variation of a DP Mechanism [36]) *A mechanism $\mathcal{M}$ has total variation at most $TV(\mathcal{M})$ if, for all neighboring datasets D, D',*

$$\sup_{S \subseteq \Theta} |\Pr(\mathcal{M}(D) \in S) - \Pr(\mathcal{M}(D') \in S)| \leq TV(\mathcal{M}).$$

For any (ε, δ)-DP mechanism, the TV distance can be bounded [50] as

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

This bound is generally tight, although for some DP mechanisms (e.g., Generalized Randomized Response) the exact TV can be computed more precisely [36].

3.1.1. DP Mechanisms

DP is a mathematical notion that provides formal guarantees. It can be practically implemented with a variety of mechanisms that have been proposed in the literature [28]. We present here the relevant ones for this work.

Generalized Randomized Response: One of the simplest mechanisms that achieves DP is *Generalized Randomized Response (GRR)* [49]. Specifically, GRR achieves ε-LDP as the user applies the perturbation directly and only the perturbed values are collected. The fundamental idea of this mechanisms is derived from a surveying technique used in the domain of psychology: When asked a sensitive question, survey participants would tell the truth with probability p or lie with probability $1 - p$ which gave them plausible deniability for any given answer [91]. Adapted in the context of DP, the GRR mechanism takes a value z from data domain $\mathcal{Z}$ of size $m = |\mathcal{Z}|$. It returns the true value z with probability p and any other value from $\mathcal{Z} \setminus \{z\}$ with probability $1 - p$ [6]. For any perturbed value $z' \in \mathcal{Z}$ the following holds:

$$\Pr[\mathcal{M}_{GRR}(z) = z'] = \begin{cases} p = \frac{e^{\varepsilon}}{e^{\varepsilon}+m-1} & \text{if } z = z', \\ q = \frac{1}{e^{\varepsilon}+m-1} & \text{else.} \end{cases} \tag{3.1}$$

Note that from Equation (3.1) it follows that $\Pr[z = z'] > \Pr[z = z'']$ for any $z'' \in \mathcal{Z} \backslash \{z'\}$. The perturbed value that GRR returns is thus more likely to be the true value than to be any other value. Furthermore, we can estimate the output probability of GRR, as shown in Theorem 1.

Theorem 1 (Estimation of GRR output distribution [90]) *Given the a database $D \in \mathcal{Z}^n$, $O(D) = \{\mathcal{M}_{\text{GRR}}(z)\}_{z \in D}$, and $c(z) = |\{z \in O(D)\}|$, the output distribution estimation is for all $z \in \mathcal{Z}$:*

$$\tilde{\pi}(z) = \frac{c(z) - nq}{n(p - q)}.$$

This theorem allows us to obtain an estimation $\tilde{\pi}$ of the true data distribution π given the observed perturbed data [90]. It ensures that $\widetilde{\pi}$ is the statistically most plausible estimate of the underlying distribution given the noisy outputs. We will exploit this in later attack settings.

Exponential Mechanism: Another fundamental mechanisms that achieves DP is the *Exponential Mechanism* (EM) [65]. Unlike perturbation-based mechanisms such as GRR which add noise to data values, the EM is based on selecting an outcome $\theta \in \Theta$ based on a scoring function $u : \mathcal{Z}^n \times \Theta \to \mathbb{R}$. The function defines which output θ is "best" for the given database by giving it a higher score. The mechanism $\mathcal{M}_{Exp} : \mathcal{Z}^n \to \Theta$ selects an output θ for any database $D \in \mathcal{Z}^n$ with probability proportional to

$$\exp\left(\frac{\varepsilon \cdot u(D, \theta)}{2\Delta_u}\right),$$

where Δ_u is the *sensitivity* of the scoring function. It quantifies the maximum change in the score when a single record in the database is modified. Formally, for any databases $D, D' \in \mathcal{Z}^n$ that differ only in one record and any outcome $\theta \in \Theta$, it holds that [16]

$$|u(D, \theta) - u(D', \theta)| \leq \Delta_u. \tag{3.2}$$

A higher Δ_u indicates that small changes in the database can lead to significant variations in the score, requiring a higher level of noise to achieve DP. Conversely, a lower Δ_u means the function is more stable, allowing for less noise.

By ensuring that even lower-utility outputs have a non-zero probability of being selected, the EM yields strong privacy guarantees. Specifically, it provides $(2\varepsilon\Delta_u)$-DP [65]. Moreover, the EM can be implemented in a LDP setting, where each user individually applies the mechanism to their own data before sharing it.

Gaussian Mechanism: A widely used perturbation-based mechanism is the *Gaussian Mechanism* (GM) [28]. The GM achieves DP by directly adding random noise drawn from a Gaussian distribution to the output of a function $g : \mathcal{Z}^n \to \mathbb{R}^d$. Specifically, given a database $D \in \mathcal{Z}^n$, the mechanism is defined as

$$\mathcal{M}_{GM}(D) = g(D) + \mathcal{N}(0, \sigma^2 I_d),$$

where σ is the noise scale and I_d is the d-dimensional identity matrix. How much noise is needed is defined by the ℓ_2-*sensitivity* of g, defined as

$$\Delta_g = \max_{D,D'} \|g(D) - g(D')\|_2, \tag{3.3}$$

where the maximum is taken over all pairs of neighboring databases D, D'.

The GM achieves approximate DP if the noise is calibrated to the sensitivity according to

$$\sigma \geq \frac{\Delta_g \sqrt{2 \ln(1.25/\delta)}}{\varepsilon},$$

for $0 < \varepsilon < 1$ [28]. It plays a central role in privacy-preserving ML, as it serves as the building block for the Differentially Private Stochastic Gradient Descent (DP-SGD) [1] algorithm, which we will evaluate in later attack settings.

For the GM, we have a closed-form solution for both its TV and trade-off function [24]:

$$\mathrm{TV}(\mathcal{M}_{GM}) = 2\,\Phi\Big(\frac{\Delta_g}{2\sigma}\Big) - 1,$$

where Φ is the standard normal cumulative distribution function. The corresponding f-function is [24]:

$$f_{\mathcal{M}_{GM}}(\alpha) = \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{\Delta_g}{\sigma}\Big).$$

3.2. Data Reconstruction

While the literature on privacy attacks against DP has largely focused on membership inference, data reconstruction attacks provide a more general framework. Data reconstruction can model both membership inference and attribute inference as special cases. Following Balle et al., we adopt this unified perspective and treat both membership and attribute inference as a specific instance of data reconstruction attacks [10].

In data reconstruction, the adversary targets a record z from database D with the goal of inferring any sensitive information on z. Data reconstruction attacks can also model attribute inference by considering that each record $z \in \mathcal{Z}$ is a tuple $z = (\mathbf{x}, s)$ with public features $\mathbf{x} \in \mathcal{X}^r$ and sensitive attribute $s \in \mathcal{S}$. The adversary has access to $\mathbf{x}$ for their target z and seeks to infer the sensitive attribute s.

Adversaries can be categorized by the amount of knowledge that they have to carry out the attack. We distinguish general knowledge of the adversary, e.g. about the data distribution π or about other non-target records $\hat{z} \in \mathcal{Z}$, from knowledge that is specific to the target z of the attack. We refer to the former as background knowledge and to the latter as target-specific auxiliary knowledge *aux*.

Based on their background knowledge and *aux*, an adversary can perform a reconstruction attack. Formally, such a reconstruction attack is defined as:

Definition 6 (Data reconstruction attack [10]) *Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a DP mechanism, D_- a fixed dataset and $z \in D$ an arbitrary record. Let furthermore $\ell\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ be an error function and $a(z)$ the target-specific auxiliary knowledge pertaining to z. Additionally, let $D = D_- \cup \{z\}$. Then, the reconstruction attack A proceeds as follows*

1. $z' \leftarrow A(\mathcal{M}(D); a(z))$.

2. *Output* $\ell(z, z')$.

The adversary uses their access to the DP mechanism's output, their background knowledge as well as optional target-specific knowledge *aux* to create a reconstruction candidate z'. The attack then evaluates how accurate the reconstruction of z is based on the error function ℓ. Which error function to use is highly dependent on the context. For instance, in location inference tasks ℓ could be the Euclidean distance between the predicted and the true location [10].

A reconstruction attack is successful if an adversary achieves a low reconstruction error $\ell(z, z')$ with a high probability. This intuition is captured by the performance metric for data reconstruction attacks: ReRo.

3.2.1. Reconstruction Robustness

Intuitively, ReRo states that an adversary can only achieve a reasonably good reconstruction of their target with a certain probability γ. The parameter η controls what a reasonably good reconstruction is: It bounds the value of the reconstruction error ℓ between the original record z and the adversary's reconstruction $A(\theta)$ based on the output of the DP mechanism $\mathcal{M}$. If $\eta = 0$, we speak of *perfect reconstruction*, for $\eta \geq 0$ of *partial reconstruction*. Formally, ReRo is defined as follows:

Definition 7 (Reconstruction Robustness [10]) *Let π be a prior over $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ a reconstruction error function. Mechanism $\mathcal{M} : \mathcal{Z}^n \to \Theta$ is (η, γ)-reconstruction robust with respect to π, ℓ if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction adversary $A : \Theta \to \mathcal{Z}$ it holds that*

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

It is important to note that, while reconstruction attacks allow for target-specific auxiliary knowledge *aux* (cf. Definition 6), the definition of ReRo introduced by Balle et al. does not include such knowledge. Consequently, ReRo can only be applied to attacks that assume no *aux*. The same limitation carries over to any bound or metric derived from it.

Balle et al. prove that DP implies ReRo, as stated in Theorem 2. The degree of protection against data reconstruction is parametrized both by the privacy budget ε and the *upper baseline error* $\kappa^+_{\pi,\eta}$ that is defined as:

$$\kappa^+_{\pi,\eta} = \sup_{z' \in \mathcal{Z}} \Pr_{Z \sim \pi}[\ell(Z, z') \leq \eta]. \tag{3.4}$$

The upper baseline error describes the maximum probability of an adversary to succeed without exploiting their access to the output of the DP mechanism [10].

Theorem 2 (ε-DP implies ReRo [10]) *Let π be a prior, $\eta > 0$ and $\kappa^+ = \kappa^+_{\pi,\eta}$. If a mechanism $\mathcal{M}$ satisfies ε-DP, then it also satisfies (η, γ)-ReRo with $\gamma = \kappa^+ e^{\varepsilon}$.*

More recently, Hayes et al. prove a novel bound on ReRo that is tightly related to the trade-off function in f-DP (cf. Definition 4) [42]. For $Q_{D_-} = \mathcal{M}(D_-)$ denoting the output distribution of the mechanism on the fixed dataset D_- and $P_{D_z} = \mathcal{M}(D_- \cup \{z\})$ for any target z, they give the following bound:

Theorem 3 (Novel ReRo bound [42]) *Suppose that for every dataset D_- there exists a pair of distributions $P^*_{D_-}, Q^*_{D_-}$ such that*

$$\sup_{z \in \mathcal{Z}} 1 - T(P_{D_z}, Q_{D_-})(\alpha) \leq 1 - T(P^*_{D_-}, Q^*_{D_-})(\alpha)$$

for all $\alpha \in [0, 1]$. Then $\mathcal{M}$ is (η, γ)-ReRo with

$$\gamma = \sup_{D_-} 1 - T(P^*_{D_-}, Q^*_{D_-})(\kappa^+_{\pi,\eta}).$$

Intuitively, this result shows that the adversary's probability of successfully reconstructing the target can be controlled by the trade-off function evaluated on a suitably chosen pair of distributions, in the worst case over all target points.

While ReRo provides a convenient way to quantify an adversary's ability to reconstruct individual records, it suffers from a key limitation: it does not distinguish whether the reconstruction success arises from information leaked by the mechanism $\mathcal{M}$, or from the adversary's background knowledge or population-level patterns. This limitation leads to what is known as the *privacy fallacy* [17], where privacy leakage is potentially overestimated.

3.2.2. Unbiased Reconstruction Robustness

To overcome this issue, Guerra-Balboa et al. proposed U-ReRo, a metric that explicitly corrects for such biases and is claimed to provide a more accurate assessment of privacy leakage [37]. U-ReRo builds upon ReRo but introduces a correction term to correct for the overestimation effect. It is defined as follows:

Definition 8 (Unbiased Reconstruction Robustness [37]) *Let π be a prior over $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function. Mechanism $\mathcal{M} : \mathcal{Z}^n \to \Theta$ is (η, γ)-unbiased reconstruction robust with respect to π, l if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction adversary $A : \Theta \to \mathcal{Z}$ it holds that*

$$\Pr_{Z \sim \pi, \theta \sim \mathcal{M}(D_- \cup \{Z\})}[\ell(Z, A(\theta)) \leq \eta] - \mathbb{E}_{Z_0 \sim \pi}[\Pr_{Z_1 \sim \pi, \theta \sim \mathcal{M}(D_- \cup \{Z_0\})}[\ell(Z_1, A(\theta)) \leq \eta]] \leq \gamma.$$

U-ReRo compares the probability of correctly reconstructing a target when it is a record from the database to the probability of reconstructing it when it is drawn from the underlying data distribution π. Like its predecessor metric ReRo, U-ReRo supports evaluation of both perfect and partial reconstruction through its parameter η.

Guerra-Balboa et al. show that DP also implies U-ReRo:

Theorem 4 ((ε, δ)-DP implies (η, γ)-U-ReRo [37]) *If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, then it also satisfies (η, γ)-U-ReRo with*

$$\gamma = \min\{\kappa^+_{\pi,\eta}(e^\varepsilon - 1) + \delta m, \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\}.$$

They also provide a tighter result for the case of perfect reconstruction ($\eta = 0$).

Theorem 5 ((ε, δ)-DP implies $(0, \gamma)$-U-ReRo [37]) *Let $\kappa_0^+ = \kappa_{\pi,0}^+$ and $m = |\mathcal{Z}|$. If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, then it also satisfies $(0, \gamma)$-U-ReRo with*

$$\gamma = \min\{\kappa_0^+(\mathrm{e}^\varepsilon - 1) + \delta m, \kappa_0^+(m-1)\frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1} + \kappa_0^+ - \kappa_0^-\},$$

where $\kappa_0^- = \inf_{z' \in \mathcal{Z}} \Pr_{Z \sim \pi}[\ell(z', Z) = 0]$.

4. Improving U-ReRo

In this chapter, we revisit the notion of U-ReRo and address its limitations. Existing results provide bounds on U-ReRo under the assumption that the adversary possesses no target-specific auxiliary knowledge *aux*. As a first step strengthen the bounds in this restricted setting.

The assumption of no target-specific auxiliary knowledge is a strong one: in many realistic attack scenarios, such as attribute inference, the adversary already has partial information about their target, e.g. [97, 32, 86, 71]. In contrast to the claims of Balle et al. and Guerra-Balboa et al., neither the original definitions of ReRo and U-ReRo nor their associated bounds apply to this more general and practically relevant case. To address this gap, we introduce *Aux-Aware U-ReRo*, a generalization of U-ReRo that accounts for adversaries with target-specific auxiliary knowledge, thereby yielding a truly universal metric for reconstruction risk.

4.1. Novel Bounds on Classic U-ReRo

We provide new bounds for U-ReRo that improve over the early bounds provided by [37]. Note that these bounds, following the original definition of U-ReRo and only apply to settings without target-specific auxiliary knowledge.

For these theoretical bounds, we introduce κ_π as the *average base line error*, defined in Equation (4.1).

$$\kappa_\pi = \sum_{z \in \mathcal{Z}} \Pr_{Z \sim \pi}[Z = z]\pi(z) = \sum_{z \in \mathcal{Z}} \pi(z)^2. \tag{4.1}$$

It expresses the probability of the adversary obtaining a perfectly accurate reconstruction by sampling both the reconstruction target and the reconstruction candidate according to the data distribution π. This captures a baseline probability of obtaining a correct reconstruction for any data distribution.

First, we present a general improved bound that applies both to perfect and partial reconstruction:

Theorem 6 ((ε, δ)-DP implies (η, γ)-U-ReRo) *Let π and $\eta \geq 0$ follow Definition 8, κ_π follow Equation* (4.1) *and $\kappa^+_{\pi,\eta}$ follow Equation* (3.4)*. If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, then it also satisfies (η, γ)-U-ReRo with*

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

If $\delta = 0$, then

$$\gamma \leq \min\{TV(\mathcal{M})(1-\kappa_\pi), \kappa^+_{\pi,\eta}(\mathrm{e}^\varepsilon - 1)\} \leq \min\{\frac{e^\varepsilon - 1}{e^\varepsilon + 1}(1-\kappa_\pi), \kappa^+_{\pi,\eta}(\mathrm{e}^\varepsilon - 1)\}.$$

Proof 1 *We denote by $S_\mu(z_1) = \{\theta \in \Theta \colon l(z_1, \theta) \leq \eta)\}$. Moreover, note that by the post-processing property of DP, $\mathcal{A} = A \circ \mathcal{M}$ is also an (ε, δ)-DP mechanism, moreover, if $\mathcal{M}$ is f-DP, then $\mathcal{A}$ is f-DP and $TV(\mathcal{A}) \leq TV(\mathcal{M})$([24], Theorem 2). Therefore, we proof the fist bound since*

$$\begin{aligned}
&\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_Z)}}[l(Z, A(\theta)) \leq \eta] - \mathbb{E}_{Z_0\sim\pi}\left[\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_0})}} l(Z, A(\theta)) \leq \eta\right] \\
&= \sum_{z_1\in\mathcal{Z}} \Pr_{\theta\sim\mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) \leq \eta]\pi(z_1) - \sum_{z_0\in\mathcal{Z}}\sum_{z_1\in\mathcal{Z}} \Pr_{\theta\sim\mathcal{M}(D_{z_0})}[l(z_1, A(\theta)) \leq \eta]\pi(z_1)\pi(z_0) \\
&= \sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}(D_{z_1}) \in S_\eta(z_1)]\pi(z_1) - \sum_{z_0\in\mathcal{Z}}\sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}(D_{z_0}) \in S_\eta(z_1)]\pi(z_1)\pi(z_0) \\
&\overset{(*)}{=} \sum_{z_0\in\mathcal{Z}}\sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}(D_{z_1}) \in S_\eta(z_1)]\pi(z_1)\pi(z_0) - \sum_{z_0\in\mathcal{Z}}\sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}(D_{z_0}) \in S_\eta(z_1)]\pi(z_1)\pi(z_0) \\
&= \sum_{z_0,z_1} \left(\Pr[\mathcal{A}(D_{z_1}) \in S_\eta(z_1)] - \Pr[\mathcal{A}(D_{z_0}) \in S_\eta(z_1)]\right)\pi(z_1)\pi(z_0) \\
&\overset{(**)}{=} \sum_{z_0} \pi(z_0) \sum_{z_1\neq z_0} \left(\Pr[\mathcal{A}(D_{z_1}) \in S_\eta(z_1)] - \Pr[\mathcal{A}(D_{z_0}) \in S_\eta(z_1)]\right)\pi(z_1) \\
&\overset{Def.5}{\leq} \sum_{z_0} \pi(z_0) \sum_{z_1\neq z_0} TV(\mathcal{A})\pi(z_1) \\
&\overset{[24]}{\leq} \sum_{z_0} \pi(z_0) \sum_{z_1\neq z_0} TV(\mathcal{M})\pi(z_1) \\
&= TV(\mathcal{M}) \sum_{z_0} \pi(z_0) \sum_{z\neq z_0} \pi(z) \\
&= TV(\mathcal{M}) \sum_{z_0} \pi(z_0)(1-\pi(z_0)) \\
&= TV(\mathcal{M})(1 - \sum_z \pi(z)^2).
\end{aligned}$$

where $(*)$ follows since $\sum_{z_0} \pi(z_0) = 1$ and $(**)$ holds trivially since for $z_1 = z_0$

$$\Pr_{\theta\sim\mathcal{M}(D_{z_1})}[A(\theta) = z_1] = \Pr_{\theta\sim\mathcal{M}(D_{z_0})}[A(\theta) = z_1] = \Pr_{\theta\sim\mathcal{M}(D_{z_0})}[A(\theta) = z_0]. \tag{4.2}$$

Therefore the subtraction is zero.

For the second bound we use the already known Theorem 2 [10] for pure DP. We denote:

- $A \equiv \Pr_{\substack{z\sim\pi \\ \theta\sim\mathcal{M}(D_z)}}[l(z, A(\theta)) \leq \eta]$,

- $B \equiv \mathbb{E}_{Z_0\sim\pi}\left[\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_0})}} l(Z, A(\theta)) \leq \eta\right]$.

Using this notation, by definition of (η,γ)-U-ReRo, we see that $\gamma = A - B \leq \min\{\kappa_\eta(\mathrm{e}^\varepsilon - 1), \frac{\mathrm{e}^\varepsilon-1}{\mathrm{e}^\varepsilon+1}\}$.

$$
\begin{aligned}
B &\equiv \mathbb{E}_{Z_0\sim\pi} \Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_{Z_0})}} [l(Z, A(\theta)) \leq \eta] \\
&= \sum_{z_0\in\mathcal{Z}} \Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_{z_0})}} [l(Z, A(\theta)) \leq \eta]\, \pi(z_0) \\
&\overset{(***)}{\geq} \mathrm{e}^{-\varepsilon} \sum_{z_0\in\mathcal{Z}} \Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_Z)}} [l(Z, A(\theta)) \leq \eta]\, \pi(z_0) \\
&= \mathrm{e}^{-\varepsilon} A \sum_{z_0\in\mathcal{Z}} \pi(z_0) - \delta m = \mathrm{e}^{-\varepsilon} A.
\end{aligned}
$$

where $(***)$ *follows since*

$$
\begin{aligned}
&\Pr_{\theta\sim\mathcal{M}(D_{z_1})} [l(z_1, A(\theta)) \leq \eta] = \sum_{O\in\Theta} \mathbf{1}_{\{l(z_1,A(O))\leq\eta\}} \Pr(\mathcal{M}(D_{z_1}) = O) \\
&\leq \sum_{O\in\Theta} \mathbf{1}_{\{l(z_1,A(O))\leq\eta\}} \mathrm{e}^\varepsilon \Pr(\mathcal{M}(D_{z_2}) = O) = \mathrm{e}^\varepsilon \sum_{O\in\Theta} \mathbf{1}_{\{l(z_1,A(O))\leq\eta\}} \Pr(\mathcal{M}(D_{z_2}) = O) \\
&= \mathrm{e}^\varepsilon \Pr_{\theta\sim\mathcal{M}(D_{z_2})} [l(z_1, A(\theta)) \leq \eta]
\end{aligned}
$$

In addition, Theorem 2 [10] states:

$$
A \equiv \Pr_{\substack{z\sim\pi \\ \theta\sim\mathcal{M}(D_z)}} [l(z, A(\theta)) \leq \eta] \leq \kappa_{\pi,l}(\eta)\mathrm{e}^\varepsilon \equiv \kappa_\eta \mathrm{e}^\varepsilon
$$

Therefore, aggregating both results we obtain:

$$
A - B \leq A - \mathrm{e}^{-\varepsilon} A = A(1 - \mathrm{e}^{-\varepsilon}) \leq \kappa_{\pi,l}(\eta)\mathrm{e}^\varepsilon(1 - \mathrm{e}^{-\varepsilon}) = \kappa_{\pi,l}(\eta)(\mathrm{e}^\varepsilon - 1). \tag{4.3}
$$

The right-hand side of the theorem provides a general upper bound on U-ReRo for any (ε,δ)-DP mechanism. If the specific DP mechanism is known, one can compute its TV more precisely to obtain a tighter bound, as per the left-hand side.

Next, we present a set of bounds that are only applicable to perfect reconstruction, i.e. $\ell(z, A(\theta)) = \eta = 0$. This case is particularly relevant in practice, since many privacy risks hinge on exact recovery of sensitive data. For example, in attribute inference, many sensitive attributes such as disease or religious belief are categorical in nature and do not trivially support partial reconstruction, e.g. [32, 33].

We prove two sets of bounds: First, we give bounds for U-ReRo and perfect reconstruction under (ε,δ)-DP. These are *black-box* bounds insofar as they do not require knowledge of the DP mechanism. The second set of bounds are *white-box* bounds under f-DP. For these, knowledge of the specific DP mechanism is required to compute its TV and trade-off function.

We require the following lemma to prove the novel bounds on U-ReRo under perfect reconstruction:

Lemma 1 *Given $|\mathcal{Z}| = m$, $D_z = D_- \cup \{z\}$ and $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ an (ε, δ)-DP mechanism, and $A\colon \Theta \to \mathcal{Z}$, we have that*

$$\Gamma = \sum_y \sum_{x \neq y} p_{\mathcal{A}}[z \mid D_z] - p_{\mathcal{A}}[z_0 \mid D_{z_0}] \leq m(m-1)\frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}.$$

Proof 2 *We use the following notational shortcuts*

$$\alpha_z \;=\; p_{\mathcal{A}}(z \mid D_z), \quad \beta_z^{z'} \;=\; p_{\mathcal{A}}(z \mid D_{z'}), \quad \textit{for all } z \neq z'.$$

and $\Gamma_{z'}^z = \alpha_z - \beta_{z'}^z$, together with the observation that for all $z \in \mathcal{Z}$, $\mathcal{A}$ is (ε, δ)-DP due to post-processing [24], we obtain that

$$\Gamma_z^{z'} = \alpha_z - \beta_{z'}^z \leq \beta_{z'}^z(e^\varepsilon - 1) + \delta \Leftrightarrow \beta_{z'}^z \geq \frac{\Gamma_z^{z'} - \delta}{e^\varepsilon - 1}. \tag{4.4}$$

By definition of the probability mass function (PMF):

$$\begin{aligned} p_{\mathcal{A}}(z \mid D_z) + \sum_{z' \in \mathcal{Z} \setminus \{z\}} p_{\mathcal{A}}(z' \mid D_z) &= 1 \\ \overset{(*)}{\Leftrightarrow} \alpha_z + \sum_{z' \neq z} \beta_{z'}^z &= 1 \Leftrightarrow \\ \beta_{z''}^z + \Gamma_{z''}^z + \sum_{z' \neq z} \beta_{z'}^z &= 1 \textit{ for any } z'' \neq z. \end{aligned}$$

Now combining previous equations we get:

$$\begin{aligned} \sum_{z \in \mathcal{Z}} \sum_{z'' \neq z} \Gamma_{z''}^z + \beta_{z''}^z + (m-1) \sum_{z' \neq z} \beta_z^{z'} &= m \Leftrightarrow \\ \sum_{z \in \mathcal{Z}} \sum_{z'' \neq z} \beta_{z''}^z + \sum_{z' \neq z} \beta_z^{z'} &= m - \Gamma \Leftrightarrow \\ m(m-1) \sum_{z \in \mathcal{Z}} \sum_{z'' \neq z} \beta_{z''}^z &= m - \Gamma \Leftrightarrow \\ \frac{m}{(e^\varepsilon - 1)}(\Gamma - 1\delta) &\leq m - \Gamma \Leftrightarrow \\ \Gamma(m + e^\varepsilon - 1) &\leq (e^\varepsilon - 1 + m\delta) \Leftrightarrow \\ \Gamma &\leq m(m-1)\frac{e^\varepsilon - 1 + m\delta}{m + e^\varepsilon - 1}. \end{aligned}$$

Using the previous lemma we can directly prove a bound for when the trade-off function is unknown– for instance in a black-box auditing scenario. We then use the (ε, δ)-based bounds as enunciated in the following theorem:

Theorem 7 (Bound for $(0, \gamma)$-U-ReRo under (ε, δ)-DP for perfect reconstruction) *If $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is f-DP, then for any attack $A\colon \Theta \to \mathcal{Z}$*

$$\gamma \leq 2A \sum_{s=1}^{\lfloor k/2 \rfloor} \pi_{i_s} \pi_{j_s} + (B + (k \bmod 2) \cdot A)\pi_{i_r} \pi_{j_r}$$

where (i_s, j_s) *are the* $\lfloor k/2 \rfloor$ *different pairs with higher product value* $\pi_i \pi_j$, (i_r, j_r) *the following bigger pair, i.e.,* $\lfloor k/2 \rfloor + 1$*) and A, B, D defined as:*

$$A = \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 2},$$

$$D = \frac{e^\varepsilon - 1 + m\delta}{e^\varepsilon + m - 1}, \quad k = \left\lfloor \frac{m(m-1)D}{A} \right\rfloor, \quad B = m(m-1)D - kA.$$

As an instance of the previous theorem we obtain a bound for the specific case of a uniform data distribution $\pi = U[m]$:

Corollary 1 ($(0, \gamma)$**-U-ReRo under uniform prior and** (ε, δ)**-DP)** *Let* $\pi = \mathrm{U}[m]$ *the uniform distribution over* $\mathcal{Z}$ *and* $m = |\mathcal{Z}|$*. If a mechanism* $\mathcal{M}$ *satisfies* (ε, δ)*-DP, then it also satisfies* $(0, \gamma)$*-U-ReRo with*

$$\gamma \leq \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1} \frac{m-1}{m}.$$

Proof 3 *It follows directly from Theorem 7 since all* $\pi_i = \frac{1}{m}$*.*

While previous bounds use the (ε, δ)-DP definition, we can also obtain bounds in terms of f-DP. For that we first prove the following lemmas:

Lemma 2 (Existence of a symmetric optimizer) *For any feasible* (ε, δ)*-DP mechanism* $\mathcal{M}$*, there exists a feasible mechanism* $\overline{\mathcal{M}}$ *whose diagonal entries are all equal to the same value p and whose off-diagonal entries are all equal to the same value q, and which satisfies*

$$S(\mathcal{M}) = \sum_z \left(1 - \mathcal{M}_{z,z}\right) = n - \sum_z \mathcal{M}_{z,z} = S(\overline{\mathcal{M}}).$$

In particular, an optimizer of S may be chosen of this two-parameter form.

Proof 4 *Let* S_n *be the group of permutations of* $\{1, \ldots, n\}$*. For a permutation* $\pi \in S_n$ *let* P_π *denote the corresponding permutation matrix. Define the relabeled mechanism*

$$T_\pi(\mathcal{M}) := P_\pi^\top \, \mathcal{M} \, P_\pi.$$

Conjugation by P_π *permutes both inputs and outputs by the same permutation:* $(T_\pi(M))_{z',z} = \mathcal{M}_{\pi(z'),\pi(z)}$*. Two facts are immediate:*

1. Feasibility preserved: *If* $\mathcal{M}$ *satisfies* (ε, δ)*-LDP then so does* $T_\pi(\mathcal{M})$*, because the LDP inequalities are invariant under simultaneous relabeling of inputs and outputs.*

2. Objective preserved: *The diagonal sum is invariant under conjugation,*

$$\sum_z (T_\pi(\mathcal{M}))_{z,z} = \sum_z \mathcal{M}_{z,z},$$

so $S(T_\pi(\mathcal{M})) = S(\mathcal{M})$*.*

Now average $\mathcal{M}$ over all permutations:

$$\overline{\mathcal{M}} := \frac{1}{m!} \sum_{\pi \in S_n} T_\pi(\mathcal{M}).$$

Because the set of ε-LDP channels is convex, $\overline{\mathcal{M}}$ is feasible. Linearity of S implies

$$S(\overline{\mathcal{M}}) = \frac{1}{m!} \sum_{\pi} S(T_\pi(\mathcal{M})) = S(\mathcal{M}).$$

By the action of S_n on matrix indices, all diagonal entries of $\overline{\mathcal{M}}$ are equal (they are in one orbit) and all off-diagonal entries are equal (they form another orbit). Hence there exist numbers p and q with

$$\overline{\mathcal{M}}_{z,z} = p \quad \forall z, \qquad \overline{\mathcal{M}}_{z',z} = q \quad \forall z' \neq z,$$

and from column normalization $p + (n-1)q = 1$. This completes the proof.

Lemma 3 (Bound in terms of f) *Given $|\mathcal{Z}| = m$, an (ε, δ)-DP mechanism $\mathcal{M} \colon \mathcal{Z}^n \to \Theta$, and $A \colon \Theta \to \mathcal{Z}$, we have that*

$$\Gamma \leq (m-1)m \max_{\alpha \in [a,b]} (1 - f(\alpha) - \alpha)$$

with $a = \frac{1-\delta}{m-1+\mathrm{e}^{\varepsilon}}$, $b = \frac{1+\delta \mathrm{e}^{-\varepsilon}}{m-1+\mathrm{e}^{-\varepsilon}}$.

Proof 5

$$\begin{aligned}
\Gamma &= \sum_z \sum_{z' \neq z} p_{\mathcal{A}}[z \mid D_z] - p_{\mathcal{A}}[z \mid D_{z'}] \\
&\overset{[56]}{\leq} \sum_z \sum_{z' \neq z} 1 - f(p_{\mathcal{A}}[z \mid D_{z'}]) - p_{\mathcal{A}}[z \mid D_{z'}] \\
&= m(m-1) - \sum_z \sum_{z' \neq z} f(p_{\mathcal{A}}[z \mid D_{z'}]) - p_{\mathcal{A}}[z \mid D_{z'}] \\
&\leq m(m-1)\left(1 - f\left(\alpha\right) - \alpha\right),
\end{aligned}$$

where

$$\alpha = \frac{\sum_z \sum_{z' \neq z} p_{\mathcal{A}_z}[z \mid D_{z'}]}{m(m-1)}.$$

So it is only remaining to find the interval where α belongs. We observe that

$$\sum_z \sum_{z' \neq z} p_{\mathcal{A}}[z \mid D_{z'}] = \sum_{z'} \sum_{z \neq z'} p_{\mathcal{A}}[z \mid D_{z'}] = \sum_z 1 - p_{\mathcal{A}}[z \mid z] = S(\mathcal{A})$$

Since D_- is fix, we can consider $\mathcal{A}$ and (ε, δ)-LDP mechanism and apply Lemma 2. Therefore, the limits of α get reduced to a two-variable optimization.

From the lemma, it follows that we may restrict attention to mechanisms of the form

$$\mathcal{A}_{z,z} = p, \qquad \mathcal{A}_{z',z} = q \quad (z' \neq z),$$

with $p, q \geq 0$ *and*

$$p + (m-1)q = 1. \tag{1}$$

hence, our objective becomes

$$S(\mathcal{A}) = S(p) = m(1-p),$$

so maximizing (resp. minimizing) S *is equivalent to minimizing (resp. maximizing)* p *the* (ε, δ)*-LDP constraints:*

$$p \leq e^{\varepsilon} q + \delta, \qquad q \leq e^{\varepsilon} p + \delta.$$

The second inequality is equivalently

$$p \geq e^{-\varepsilon}(q - \delta).$$

Combining with nonnegativity, we may write the feasible range for p *(for a given* q*) as*

$$\max\{0,\ e^{-\varepsilon}(q-\delta)\} \ \leq \ p \ \leq \ e^{\varepsilon} q + \delta.$$

Using the normalization $q = (1-p)/(m-1)$*, we can eliminate* q *and obtain explicit bounds on* p*.*

Maximum of S**.** *To maximize* $S(p) = m(1-p)$ *we minimize* p*. The minimum* p *allowed is* $p \geq e^{-\varepsilon}(q - \delta)$*. together with the normalization* $p + (m-1)q = 1$ *yields, after substituting* $q = (1-p)/(m-1)$*,*

$$p \geq e^{-\varepsilon}\left(\frac{1-p}{m-1} - \delta\right).$$

Rearranging gives

$$p\left(1 + \frac{e^{-\varepsilon}}{m-1}\right) \geq e^{-\varepsilon}\left(\frac{1}{m-1} - \delta\right),$$

hence

$$p \ \geq \ \frac{e^{-\varepsilon}\big(1 - (m-1)\delta\big)}{m - 1 + e^{-\varepsilon}} = \frac{\big(1 - (m-1)\delta\big)}{(m-1)\mathrm{e}^{\varepsilon} + 1}.$$

hence

$$S_{\max} = m\left(1 - \frac{e^{-\varepsilon}\big(1 - (m-1)\delta\big)}{m - 1 + e^{-\varepsilon}}\right) = m(m-1)\frac{\mathrm{e}^{\varepsilon} + \delta}{(m-1)\mathrm{e}^{\varepsilon} + 1}.$$

Hence,

$$\alpha_{\max} = \frac{(\mathrm{e}^{\varepsilon} + \delta)}{(m-1)e^{\varepsilon}} + 1 = \frac{(1 + \mathrm{e}^{-\varepsilon}\delta)}{(m-1) + e^{-\varepsilon}}.$$

Minimum of S. *To minimize S we maximize p . The largest p allowed is $p = e^{\varepsilon}q + \delta$. Hence,*

$$e^{\varepsilon}q + \delta + (m-1)q = 1 \quad \Longrightarrow \quad q = \frac{1-\delta}{m-1+e^{\varepsilon}}, \qquad p = \frac{e^{\varepsilon}+\delta(m-1)}{m-1+e^{\varepsilon}}.$$

Therefore the minimum value of S is

$$S_{\min} = m(1-p) = m\left(1 - \frac{e^{\varepsilon}+\delta(m-1)}{m-1+e^{\varepsilon}}\right) = \frac{m(m-1)(1-\delta)}{m-1+e^{\varepsilon}}.$$

and

$$\alpha_{\min} = \frac{(1-\delta)}{m-1+e^{\varepsilon}}.$$

We obtain the following result for U-ReRo under f-DP for perfect reconstruction:

Theorem 8 (Bound for $(0,\gamma)$-U-ReRo under f-DP) *If $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is f-DP, then for any attack $A\colon \Theta \to \mathcal{Z}$*

$$\gamma \leq 2TV \sum_{s=1}^{\lfloor k/2 \rfloor} \pi_{i_s}\pi_{j_s} + (B + (k \bmod 2)\cdot TV)\pi_{i_r}\pi_{j_r}$$

where (i_s, j_s) are the $\lfloor k/2 \rfloor$ different pairs with higher product value $\pi_i\pi_j$, (i_r, j_r) the following bigger pair, i.e., $\lfloor k/2 \rfloor + 1$) and $A = TV(\mathcal{M})$, and D, B defined as:

$$D = \max_{\alpha\in[a,b]} (1 - f(\alpha) - \alpha)), \quad k = \left\lfloor \frac{D(m-1)m}{TV} \right\rfloor, \quad B = m(m-1)D - k\cdot TV.$$

with $a = \frac{1-\delta}{m-1+e^{\varepsilon}}$, $b = \frac{1-\delta e^{-\varepsilon}}{m-1+e^{-\varepsilon}}$.

Finally, we give a refined white-box bound for perfect reconstruction under uniform prior $\pi = U[m]$.

Corollary 2 (Bound on $(0,\gamma)$-U-ReRo under uniform prior and f-DP) *Let $\pi = \mathrm{U}[m]$. If $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is f-DP, then for any attack $A\colon \Theta \to \mathcal{Z}$*

$$\gamma \leq \max_{\alpha\in[a,b]} (1 - f(\alpha) - \alpha))\frac{m-1}{m}.$$

with $a = \frac{1-\delta}{m-1+e^{\varepsilon}}$, $b = \frac{1+\delta e^{-\varepsilon}}{m-1+e^{-\varepsilon}}$.

Proof 6 *Follows by direct application of previous bound to $\pi_i = \frac{1}{m}$.*

For our white-box bounds, we need to compute the f-function and the TV for a given DP mechanism. As an exemplary mechanism, we briefly describe how to compute the TV and f-function for GRR.

Example 1 (TV and f-function of GRR) *For the ε-GRR mechanism with parameters $p = \frac{e^\varepsilon}{e^\varepsilon+m-1}$ and $q = \frac{1}{e^\varepsilon+m-1}$, we compute the TV and f-function as follows:*

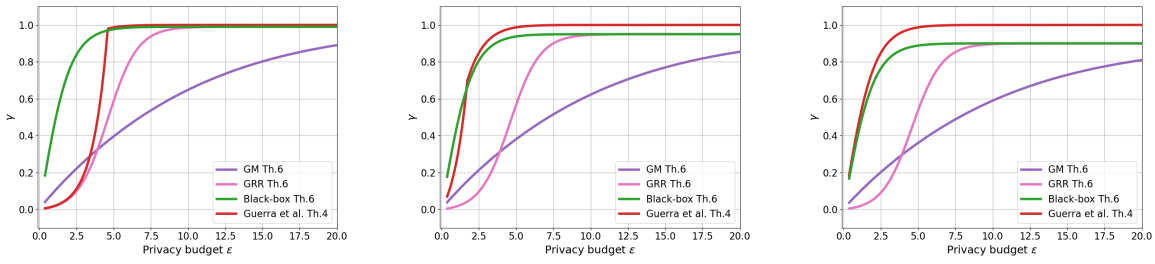
Total variation. *For two distinct inputs $z \neq z' \in \mathcal{Z}$,*

$$\mathrm{TV}\big(\mathcal{M}_{\mathrm{GRR}}(z), \mathcal{M}_{\mathrm{GRR}}(z')\big) = \tfrac{1}{2}\sum_{y\in\mathcal{Z}} \big|\Pr[y \mid z] - \Pr[y \mid z']\big| = p - q = \frac{e^\varepsilon - 1}{e^\varepsilon + m - 1}.$$

Trade-off function. *The GRR mechanism only has three types of outputs when comparing z and z': the true input z, the other input z', and all other values. The likelihood ratio $P(y)/Q(y)$ takes three values: e^ε for $y = z$, $e^{-\varepsilon}$ for $y = z'$, and 1 for the rest. The optimal test accepts $y = z$, rejects $y = z'$, and randomizes on the others. This leads to the three-piece f-function:*

$$f(\mathcal{M}_{GRR}(z), \mathcal{M}_{GRR}(z'))(\alpha) = \begin{cases} 1 - \mathrm{e}^\varepsilon\alpha & 0 \le \alpha \le q, \\ mq - \alpha & q < \alpha \le (1-p), \\ \mathrm{e}^{-\varepsilon}(1-\alpha) & (1-p) < \alpha \le 1. \end{cases}$$

Our novel bounds show a substantial improvement over the state-of-the-art bounds provided by Guerra-Balboa et al. for classic U-ReRo. As illustrated in Figure 4.1, our novel general bound (Theorem 6) consistently outperforms the earlier bound from Theorem 4. When the underlying DP mechanism is known and its TV can be computed, our bound always improves upon the baseline. Even in a black-box setting, where we rely on the worst-case TV, our bound provides an improvement whenever the data distribution is non-uniform and the baseline errors are skewed. Notably, the improvement of our approach grows with the skew in the data.

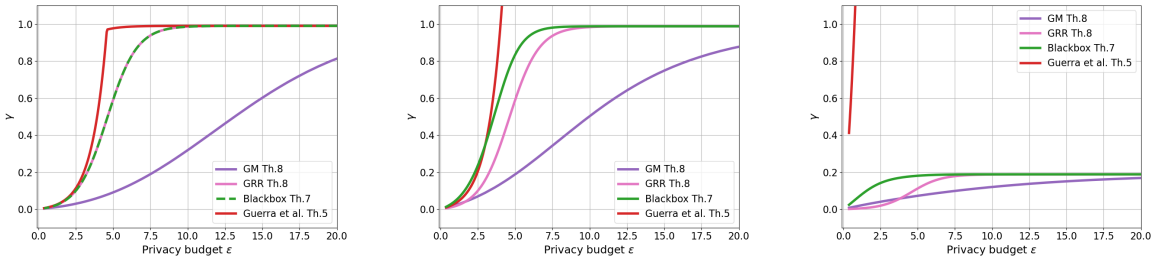


(a) $\kappa_\pi = \kappa^+_{\pi,\eta} = 0.01$. (b) $\kappa_\pi = 0.05, \kappa^+_{\pi,\eta} = 0.15$. (c) $\kappa_\pi = 0.1, \kappa^+_{\pi,\eta} = 0.4$.

Figure 4.1.: Comparison of our general bound Theorem 6 with Theorem 4 from Guerra-Balboa et al. for varying baseline errors κ_π and $\kappa^+_{\pi,\eta}$. Black-box corresponds to Theorem 6 with the general TV. Parameters $\delta = 1e^{-5}$ and $m = 100$.

In Figure 4.2, we can see that this observation extends to the setting of perfect reconstruction as well. Both our black-box bound and our bound using the f-function of specific DP mechanisms outperform the original bound from Theorem 5. The improvement becomes increasingly pronounced as the data distribution becomes more skewed towards individual records. Even under a uniform data distribution, our bounds remain tighter for $\varepsilon \geq 5$.

Finally, Figure 4.3 demonstrates that under a uniform data distribution, both our black-box bound and the f-DP bound are strictly tighter than the bound from Theorem 5. This



(a) Uniform π. (b) Skewed π. (c) Strongly skewed π.

Figure 4.2.: Comparison of our bounds for perfect reconstruction (Theorem 8 and Theorem 7) with Theorem 5 from Guerra-Balboa et al. for varying data distribution π. Parameters $\delta = 1e^{-5}$ and $m = 100$.

highlights that our improvements are consistent across different assumptions about the data distribution and the knowledge of the DP mechanism.

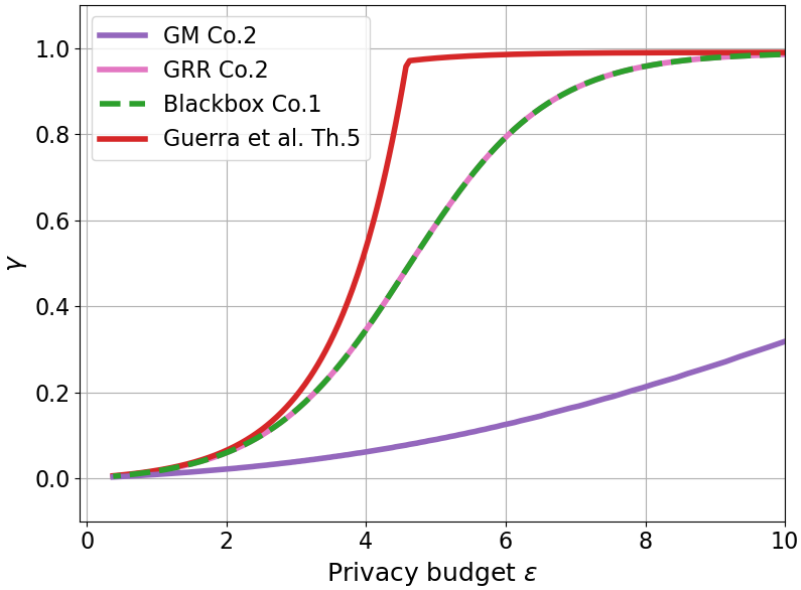


Figure 4.3.: Comparison of our bounds for perfect reconstruction and uniform prior (Corollary 2 and Corollary 1) with Theorem 5 from Guerra-Balboa et al.

Overall, our work establishes tighter and more versatile bounds for U-ReRo, with the largest gains observed in skewed data scenarios and in cases where the DP mechanism is known and its TV and f-functions can be computed, thereby significantly advancing the state of the art.

4.2. Aux-Aware U-ReRo

We now adapt classic U-ReRo to explicitly incorporate target-specific auxiliary knowledge. This new formulation, Aux-Aware U-ReRo, allows to measure the success of both mem-

bership and attribute inference attacks as special cases of data reconstruction attacks by considering varying levels of *aux*.

Definition 9 (Aux-Aware (η, γ)-U-ReRo) *Let π, ℓ follow Definition 7. For each $z \in \mathcal{Z}$, let $a(z)$ describe the target-specific auxiliary information on z. A mechanism $\mathcal{M} : \mathcal{Z}^n \to \Theta$ is aux-aware (η, γ)-unbiased reconstruction robust if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction adversary $A : \Theta \times \mathcal{X} \to \mathcal{Z}$ it holds that*

$$\Pr_{\substack{Z\sim\pi,\\ \theta\sim\mathcal{M}(D_-\cup\{Z\})}}[\ell(Z, A(\theta, a(Z))) \le \eta] - \mathbb{E}_{Z_0\sim\pi}\Big[\Pr_{\substack{Z_1\sim\pi,\\ \theta\sim\mathcal{M}(D_-\cup\{Z_0\})}}[\ell(Z_1, A(\theta, a(Z_1)) \le \eta]\Big] \le \gamma.$$

Note that the definition of Aux-Aware U-ReRo inherently includes a definition of Aux-Aware ReRo as its first term. Moreover, in the case of perfect reconstruction, data reconstruction with *aux* reduces to attribute inference. Under these conditions, Aux-Aware (η, γ)-U-ReRo recovers the classic attribute advantage introduced in [97]. Finally, we observe that when $a(z) = \varnothing$, Aux-Aware U-ReRo coincides with classic U-ReRo.

Previous bounds for ReRo and U-ReRo assume that reconstruction attacks perform equally well for every target $z \in \mathcal{Z}$. 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 more unique and unusual. Such differences are not captured by ReRo or classic U-ReRo, nor reflected in the proofs of their bounds [10, 37]. We therefore provide the first theoretical bound that explicitly accounts for *aux*.

Theorem 9 ((ε, δ)-DP implies Aux-Aware (η, γ)-U-ReRo) *Let π and $\eta \ge 0$ follow Definition 8, κ_π follow Equation (4.1) and $\kappa^+_{\pi,\eta}$ follow Equation (3.4). If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, then it also satisfies Aux-Aware (η, γ)-U-ReRo with*

$$\gamma \le TV(\mathcal{M})(1 - \kappa_\pi) \le \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1 - \kappa_\pi).$$

Proof 7 *Note that by the post-processing property of f-DP [24], given a fixed target auxiliary knowledge $a(z)$, any attack defines the f-DP mechanism*

$$\mathcal{A}(D, a(z)) = A_z(\mathcal{M}(D)),$$

and analogously for (ε, δ)-DP [28]. Therefore, we obtain

$$\begin{aligned}
&\Pr_{\substack{Z\sim\pi,\\ \theta\sim\mathcal{M}(D_-\cup\{Z\})}}[\ell(Z, A(\theta, a(Z))) \le \eta] \\
&\quad - \mathbb{E}_{Z_0\sim\pi}\Big[\Pr_{\substack{Z_1\sim\pi,\\ \theta\sim\mathcal{M}(D_-\cup\{Z_0\})}}[\ell(Z_1, A(\theta, a(Z_1))) \le \eta]\Big] \\
&= \sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}_{z_1}(D_{z_1}) \in S_\eta(z_1)]\pi(z_1) - \sum_{z_0\in\mathcal{Z}}\sum_{z_1\in\mathcal{Z}} \Pr[\mathcal{A}_{z_1}(D_{z_0}) \in S_\eta(z_1)]\pi(z_1)\pi(z_0).
\end{aligned}$$

$$\overset{(*)}{=} \sum_{z_0,z_1} \Big(\Pr[\mathcal{A}_{z_1}(D_{z_1}) \in S_\eta(z_1)] - \Pr[\mathcal{A}_{z_1}(D_{z_0}) \in S_\eta(z_1)] \Big) \pi(z_1)\pi(z_0)$$

$$\overset{(**)}{=} \sum_{z_0} \pi(z_0) \sum_{z_1 \neq z_0} \Big(\Pr[\mathcal{A}_{z_1}(D_{z_1}) \in S_\eta(z_1)] - \Pr[\mathcal{A}_{z_1}(D_{z_0}) \in S_\eta(z_1)] \Big) \pi(z_1).$$

$$\begin{aligned}
&\overset{Def.5}{\leq} \sum_{z_0} \pi(z_0) \sum_{z_1 \neq z_0} TV(\mathcal{A}_{z_1})\pi(z_1) \\
&\overset{[24]}{\leq} \sum_{z_0} \pi(z_0) \sum_{z_1 \neq z_0} TV(\mathcal{M})\pi(z_1) \\
&= TV(\mathcal{M}) \sum_{z_0} \pi(z_0)(1 - \pi(z_0)) \\
&= TV(\mathcal{M}) \Big(1 - \sum_{z} \pi(z)^2 \Big).
\end{aligned}$$

Part I.

Analyzing U-ReRo in Local Differential Privacy

5. Introduction

We first evaluate ReRo and U-ReRo in the setting of LDP–a rigorous and increasingly relevant privacy model in which data is randomized on the client side before being transmitted to a data collector [28]. Unlike the global model of DP, LDP does not rely on trust in a central server, making it especially suitable for privacy-sensitive applications such as telemetry and location-based services [30]. While LDP enjoys strong formal guarantees, it has also been shown to be vulnerable to inference attacks [6, 34, 39], making it an ideal testbed for assessing whether leakage metrics like ReRo and U-ReRo provide an accurate assessment of privacy leakage.

Our study is driven by four research questions. First, we ask whether ReRo overestimates privacy leakage in practice, thus producing overly conservative results that may mislead practitioners. Second, we investigate whether U-ReRo successfully corrects for potential overestimation and provides a more faithful assessment of the actual privacy leakage. Next, we explore how tight our theoretical bounds on leakage are in practice–that is, how closely they track the empirical leakage values observed in different attacks. Finally, we evaluate whether our novel metric, Aux-Aware U-ReRo, successfully captures privacy leakage in settings where the adversary has access to *aux*.

5.1. Location Data and the Roadgraph Model

Location data is among the most sensitive types of personal information–just a few spatio-temporal points can uniquely identify individuals [68]. It also enables valuable services such as navigation, urban planning, and epidemic tracking [80, 95]. Because of its sensitivity, utility, and real-world relevance [4, 99, 53], location data is a compelling setting for evaluating our privacy metrics.

We assume a setting where users report their current location or recent trajectory to a digital service. To protect privacy, each user perturbs their true location v using a DP mechanism $\mathcal{M}$ before sending it:

$$\widetilde{v} = \mathcal{M}(v).$$

Following the literature, we represent the environment as a directed graph $G = (V, E)$, where nodes $V(G)$ correspond to street intersections and edges $E(G)$ to road segments [9]. This graph is obtained from OpenStreetMap (OSM) [75], and defines the discrete data domain $\mathcal{Z} = V(G)$ of possible user locations. A distribution π over $\mathcal{Z}$ characterizes how frequently each location is visited.

The adversary aims to infer a user's true location v from the noisy report $\widetilde{v}$. We evaluate inference accuracy in terms of spatial proximity between the predicted location w and the actual location v, using the shortest-path distance $d_G(w, v)$ computed via Dijkstra's

algorithm [21]. The loss is defined as:

$$\ell_G(w, A(\widetilde{v})) = \begin{cases} 0 & \text{if } w = A(\widetilde{v}) \\ d_G(w, A(\widetilde{v})) & \text{otherwise.} \end{cases} \tag{5.1}$$

We define a *trajectory* as an ordered sequence of spatio-temporal points along the graph, capturing a user's movement consistent with temporal and spatial constraints [67].

5.2. Mechanism Selection

For our evaluation, we select two widely studied DP mechanisms: GRR and the EM. These mechanisms serve as fundamental building blocks for more complex privacy-preserving schemes, e.g. [94, 90, 26].

GRR is a classic mechanism that randomly reports either the true value or a different one with probabilities linked directly to ε [29], cf. Section 3.1.1. One key advantage of GRR is that we have a known optimal attack against it [6], which enables us to rigorously test the tightness of our theoretical U-ReRo bounds.

However, GRR treats all incorrect outputs equally, which limits its ability to capture privacy risks arising from partial reconstruction. Studies have shown that partial reconstruction–in the setting of location data, identifying a user's position within a certain region–can pose significant privacy risks by revealing sensitive patterns or private information [35, 55]. Since GRR's output distribution is uniform over incorrect values, it cannot model near-miss inferences, restricting our analysis to perfect reconstruction.

To address this limitation, we consider the EM, which assigns higher probabilities to outputs that are more similar to the true input according to a scoring function. For our location-based setting, we define the scoring function as the negative hop distance $-\ell_G(v, w)$ between the target node v and any other node $w \in V(G)$ (cf. Equation (5.1)).

Based on the scoring function, the EM produces outputs closer to the original value with higher probability [65]. This proximity-aware property makes EM well-suited for studying partial reconstruction risks.

For partial reconstruction, we have to choose the reconstruction threshold parameter η which in our setting ranges from 0 to the diameter Δ_G of the graph. We evaluate $\eta \in \{0.1, 0.3, 0.5, 0.7\} \cdot \Delta_G$ respectively.

5.3. Attack Descriptions

We implement four different attacks to evaluate the behavior of U-ReRo and ReRo under varying adversarial assumptions. We describe each attack in detail in this section. An overview of the setting for each attack can be found in Table 5.1.

5.3.1. Uniform Prior Attack (UNI)

The first attack assumes a uniform data distribution $\pi = U[m]$ – meaning all reconstruction candidates are equally likely– and an adversary without background knowledge or *aux*.

Attack setting	π	Knowledge of π	*aux*	U-ReRo version
UNI	$U[m]$	–	–	Classic (Definition 8)
TRUE	π	–	–	Classic (Definition 8)
PRIOR	π	π	–	Classic (Definition 8)
EST	π	–	–	Classic (Definition 8)
CORR	π	–	Prev. node v_{t-1}	Aux-Aware (Definition 9)

Table 5.1.: Data distribution and background knowledge for each attack setting.

In this scenario, any success in reconstruction must stem solely from leakage of the DP mechanism.

The attack strategy is simple [6]: the adversary uses the target's perturbed report as their prediction. We choose this attack strategy as an interpretable baseline of comparison for ReRo and U-ReRo and because prior work showed its performance closely matches the theoretical privacy leakage under the GRR mechanism [6] which we will exploit to test the tightness of our theoretical bound. The attack is formally described in Attack 1.

Attack 1 (Uniform Prior Attack (based on [6])) *Let $\mathcal{M}$ be a DP mechanism and G a graph. Let furthermore u be the target user with true location $v \in V(G)$ and perturbed location $\widetilde{v} = \mathcal{M}(v)$. The adversary*

1. *Receives the perturbed location $\widetilde{v}$,*
2. *Outputs $\widetilde{v}$.*

Since this attack relies solely on the perturbed report, we expect only a minimal difference between ReRo and U-ReRo. The correction term in U-ReRo should be small, as the adversary has access only to the perturbed location report, which does not align with the actual target of the attack. Consequently, the success probability is expected to be very low.

5.3.2. True Distribution Attack (TRUE)

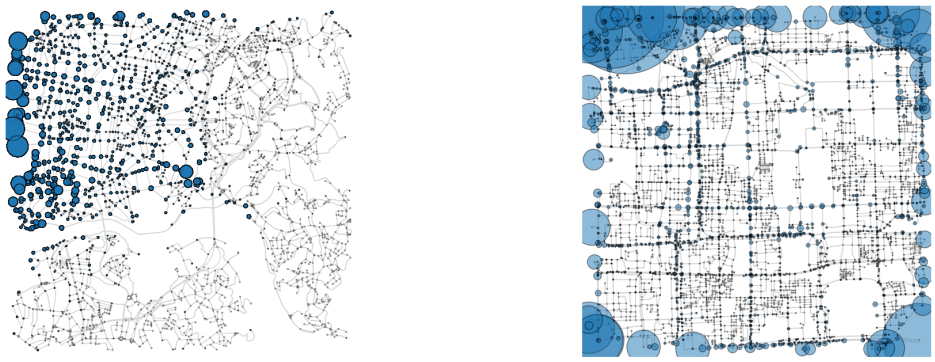
While we only have a proven optimality result for the UNI attack strategy against GRR under uniform prior, we will also evaluate the same attack against the true data distribution π. This allows us to evaluate the tightness of our bounds for non-uniform data distributions, namely Theorem 7 and Theorem 8.

5.3.3. Prior Recovery Attack (PRIOR)

The next attack models an adversary who relies exclusively on their knowledge of the data distribution π, completely ignoring the mechanism's output. Success here cannot be attributed to the DP mechanism, allowing us to examine whether ReRo conflates

adversarial knowledge with actual leakage and whether U-ReRo effectively corrects for this.

The adversary knows the data distribution π. Specifically, we base π on the fact that not all locations (nodes of graph G) are equally visited but that some are more popular than others. The knowledge about location popularity can be considered as public information or knowledge the adversary has access to due to other independent releases. We plot this distribution for our location datasets in Figure 5.1.



(a) Porto dataset.

(b) Beijing dataset.

Figure 5.1.: Data distribution π over the nodes of the Porto and Beijing datasets. Nodes are scaled according to the number of visits they receive.

The adversary retrieves the most likely node based on π and outputs it as a prediction for any target. The attack is formally described in Attack 2.

Attack 2 (Prior recovery attack) *Let G be a graph and π a probability distribution over $V(G)$. The adversary*

1. *Retrieves $w = \arg\max_{v \in V(G)} \pi(v)$, the most popular location,*
2. *Outputs w.*

Since this attack relies entirely on existing knowledge and does not incorporate any information from user reports, it does not pose a privacy risk–there is no learning of private information. Consequently, the value of U-ReRo should be zero, as the mechanism itself does not contribute to privacy leakage. However, we expect ReRo to be high when the prior is sufficiently informative, showcasing its overestimation effect.

5.3.4. Estimation-based Attack (EST)

The fourth attack explores an adversary using the output of the mechanism to estimate population-level statistics. This setup helps reveal if ReRo mistakenly interprets aggregate statistical information as privacy leakage, and whether U-ReRo properly distinguishes between the two.

In the previous case, the adversary had access to the true data distribution π. We now consider a setting where the adversary has no knowledge of the data distribution π but instead collects I outputs of the DP mechanism from randomly chosen users. Based on this collected dataset, the adversary then estimates the data distribution to obtain $\widetilde{\pi}$. As before, they use the most likely candidate, based on $\widetilde{\pi}$, as their prediction for any target.

As we have a closed-form solution for the estimation of the distribution $\widetilde{\pi}$ under the GRR mechanism (cf. Section 3.1.1, we restrict this attack to this setting only. The formal attack description is presented in Attack 3.

Attack 3 (Estimation-based Attack) *Let $\mathcal{M}_{GRR}$ be the GRR mechanism and G a graph, and let $O(D) = \{\mathcal{M}_{GRR}(v) \mid v \in D\}$ be the set of perturbed reports for database D. Given the noisy counts $c(v)$ for each $v \in V(G)$, the adversary performs the following:*

1. *Runs the GRR mechanism on I nodes, sampled according to π, obtaining noisy outputs $O(D)$.*
2. *Estimates $\tilde{\pi}(v)$ for all $v \in V(G)$ using the formula from Theorem 1:*

$$\tilde{\pi}(v) = \frac{c(v) - Iq}{I(p-q)}$$

   where p, q are GRR parameters, and $c(v)$ the number of occurrences of v in $O(D)$.
3. *Computes*

$$w = \arg\max_{v \in V(G)} \tilde{\pi}(v).$$

4. *Outputs w.*

While this approach is thematically similar to the PRIOR attack where the adversary also selects the most popular node the key difference lies in the source of knowledge. In the PRIOR attack, the adversary relies on knowledge of the true distribution π, which is independent of the mechanism's output. In contrast, the adversary in this setting learns the global statistical distribution from the output of the mechanism itself. Consequently, the attack is inherently dependent on the mechanism $\mathcal{M}$ and the privacy budget ε, as its output directly influences the accuracy of $\tilde{\pi}$.

Since the adversary does not have any knowledge of the true data distribution π, we expect the attack performance as measured by ReRo to be lower than for the PRIOR attack since the adversary obtains their predicted node based on a noisy estimate of π. U-ReRo should not be affected by this as we expect it to be close to 0 for this setting, mirroring the behavior we already saw in the PRIOR setting.

5.3.5. Correlation-based Attack (CORR)

The final attack leverages target-specific auxiliary knowledge *aux*, modeling a realistic adversary who combines both knowledge of their target and the mechanism's output. This setting enables the evaluation of our proposed Aux-Aware U-ReRo metric and its bound.

Concretely, the adversary knows the target's previous true location v_{t-1} and has access to a Markov model M over the graph G, defined by a transition matrix P that encodes the probabilities of moving between nodes. Given the high predictability of human mobility [68], the adversary uses the Markov model together with the previous location to determine whether a reported location is plausible. If it passes this plausibility check, the adversary predicts the perturbed location as the true location; otherwise, they rely solely on sampling from the Markov model. The full attack strategy is illustrated in Attack 4.

Attack 4 (Correlation-based Attack (based on [87])) *Let $\mathcal{M}$ be a DP mechanism, G a graph and $M = (V(G), P)$ a Markov model over G. For user u with true location v_t and perturbed location $\widetilde{v}_t = \mathcal{M}(v_t)$ at time t, the adversary:*

1. *Receives v_{t-1} and $\widetilde{v}_t$,*
2. *Retrieves $P(\widetilde{v}_t|v_{t-1})$,*
3. *If $P(\widetilde{v}_t|v_{t-1}) > \tau$, outputs $w = \widetilde{v}_t$,*
4. *Otherwise, outputs $w \sim P(*|v_{t-1})$.*

We expect this experiment to yield intermediate results compared to the previous attacks. Aux-Aware U-ReRo values should be higher than zero, as the adversary gains information from the user's perturbed location reports, which would not be possible without their participation in the data collection. However, they should be lower than Aux-Aware ReRo, since the adversary has access to both highly informative *aux* for their target and the Markov model and therefore should be able to predict some information regardless of whether they have access to the corresponding perturbed location report.

5.4. Database Descriptions

We work with two real-life location datasets: the Porto taxi dataset [73] and the Geolife dataset [100], referred to as Porto and Beijing datasets throughout this work. They are widely used in privacy and mobility research, e.g., [80, 58, 93]. Both datasets are also publicly available. We predominantly use the datasets to compute the data distribution π over the nodes of the graph.

The Porto dataset contains over 1.7 million taxi trips collected from 442 taxis operating in the city of Porto, Portugal, between 01/07/2013 and 30/06/2014. Each record consists of a trip ID, a starting timestamp, and a series of GPS coordinates reported every 15 seconds, forming a trajectory. We assume that each trip corresponds to a unique user, as our focus lies on global mobility patterns rather than individual mobility traces. We obtain a total of 1,710,670 trajectories, each treated as representing a distinct user.

To extract the road network, we define a circular area using a centerpoint and radius. For Porto, we chose the centerpoint at $(41.1474557, -8.5870079)$ – located in the city center – with a radius of 2688 meters. This captures the urban core where the majority of taxi activity occurs, including central districts such as Aliados and Boavista. We selected the radius to strike a balance between covering the dense urban area and excluding too

peripheral regions that introduce unnecessary road segments. It also ensures that the resulting graph remains computationally tractable. For the city of Porto, we have a total of $m = 3052$ locations.

The Beijing dataset contains GPS trajectories recorded by 182 users over a period of more than three years, from 04/2007 to 08/2012, primarily in and around Beijing, China. The dataset includes trajectories of 182 individuals, covering a wide range of transportation modes such as walking, biking, and driving. Each trajectory consists of a sequence of timestamped GPS coordinates sampled at intervals of 1–5 seconds [100]. We treat each of the 18,670 trajectories as an independent instance and assume one trajectory per user. As with the Porto dataset, we map the raw GPS coordinates to the graph representation. For the Beijing region represented in our experiments, this results in a total of $m = 5356$ locations.

For the Beijing dataset, we selected the centerpoint at $(39.9130, 116.3703)$ — approximately the location of Tiananmen Square with a radius of 5000 meters. Given the larger spatial spread of the Geolife dataset and its inclusion of diverse travel modes, this radius ensures coverage of key central districts such as Dongcheng, Xicheng and Chaoyang, where many of the trajectories are concentrated. The 5 km radius furthermore ensures that the graph remains reasonable in size and computationally tractable.

Database	N	m
Porto	3052	3052
Beijing	5356	5356

Table 5.2.: Database overview: size of universe N, number of reconstruction candidates m.

For both datasets, we compute the data distribution π over the graph nodes as follows: for each node, we count the number of visits according to the mapped locations from the full dataset; finally, we normalize these counts so that they sum to 1. This yields a probability distribution reflecting how frequently each node is visited.

For the EST attack setting, we furthermore require a fixed dataset D_{-}. We set D_{-} to 499 nodes sampled according to the data distribution π.

5.5. Experimental Design

All experiments are implemented in Python. We use the NetworkX [40] and OSMNX [15] libraries to construct directed roadgraphs for the Porto and Beijing regions based on publicly available OSM data. Each GPS location report in the dataset is mapped to its nearest graph node, resulting in a discrete data domain.

Monte Carlo approximations: For the UNI and PRIOR attacks, ReRo and U-ReRo can be computed analytically. However, in other settings such as the EST or CORR attacks, closed-form solutions are not feasible due to the complexity of the models or distributions involved. In such cases, we rely on Monte Carlo approximation methods, which are widely used for estimating expectations when analytical computation is impractical [81, 54].

We estimate these quantities using a frequentist approach by drawing I independent samples. According to the Law of Large Numbers [79], as the number of samples increases, the empirical estimates converge to the true values. To verify the quality of our approximation, we compare the empirical results with theoretical values where available. A close alignment between signifies that our chosen sample size is sufficient for reliable estimation [76].

Furthermore, to account for the inherent randomness introduced by the DP mechanism, we repeat its execution on each of the I sampled inputs J times. Following the literature [6, 22], we set $J = 5$ for our experiments, unless stated otherwise. This repetition allows us to average over multiple independent runs for the same input, thereby reducing the variance of our estimates and obtaining a more stable approximation of the underlying distribution. An overview over the choices of I and J for each attack can be found in Table 5.3.

Attack	I	J
UNI	1000	5
TRUE	1000	5
PRIOR	1000	5
EST	1000	5
CORR	1000	5

Table 5.3.: Sample sizes used for different attacks. I: number of samples, J: number of mechanism runs per sample.

ReRo: When computing ReRo, the sample $z \in \mathcal{Z}$ drawn to approximate the random variable $Z \sim \pi$ always corresponds to the target of the attack. Depending on the type of attack, we either receive the DP mechanism's output $\theta = \mathcal{M}(z)$ directly–as in the UNI or CORR attacks–or we insert the sample into an existing database D_{-}, forming $D = D_{-} \cup \{z\}$, and then query the mechanism to obtain $\theta = \mathcal{M}(D)$, as in the EST attack. In the latter case, the background dataset D_{-} remains fixed across all iterations [10]. In either case, we know that the sampled record z is included in the input to the DP mechanism.

To estimate ReRo, we perform I iterations; unless otherwise stated, we set $I = 1000$, as empirical evaluations show close alignment with theoretical values at this scale, which we demonstrate in the result section. In each iteration, a target $z \in \mathcal{Z}$ is sampled according to π, passed through the DP mechanism, and then attacked to yield a predicted value z'. The attack is considered successful if $z' = z$ in the case of perfect reconstruction (i.e., reconstruction error $\ell_G(z, z') = \eta = 0$), or if the reconstruction error corresponds to $\ell_G(z, z') \leq \eta$ in the case of partial reconstruction. To account for the probabilistic nature of the DP mechanisms, for each sampled target z, we pass it through the mechanism J times and obtain the result for this target by averaging the result across those J runs. The final ReRo value is computed as the average success rate across all I iterations.

U-ReRo: To compute U-ReRo, we subtract a correction term from the previously computed ReRo value. This correction term breaks the alignment present in the ReRo setting: here, the sample drawn as input to the DP mechanism and the target of the attack do not

necessarily correspond. We estimate this term using a nested Monte Carlo sampling procedure.

In the outer loop, we draw I records from the distribution π, just as with ReRo. This approximates the first random variable in the correction term, $Z_0 \sim \pi$ (cf. Definition 8). For each sample z_0, the adversary receives the DP mechanism's output θ and generates a prediction $z' = A(\theta)$ based on the attack strategy A. As with ReRo, we obtain the output of sample z_0 J times to account for variability in the output of the mechanism. For each output θ, we then compute the inner loop.

In the inner loop, we draw another I records from π to serve as reconstruction targets, thereby approximating $Z_1 \sim \pi$. For every pair (z_0, z_1), we evaluate whether the adversary's prediction matches the sampled target ($z' = z_1$ for perfect reconstruction) or whether the reconstruction loss satisfies $\ell_G(z', z_1) \leq \eta$ in the case of partial reconstruction.

The result is averaged first over all inner-loop targets z_1 for a given input z_0, and then across all outer-loop samples, yielding an empirical estimate of the correction term. Subtracting this value from ReRo gives the final U-ReRo score.

Upper bounds: We consider several upper bounds across different settings. For each setting, we plot only the most specific bounds—for instance, when the prior is uniform, we show only the bounds for a uniform prior, even though the more general bound from Theorem 6 also applies. Table 5.4 summarizes which bounds are relevant for each setting.

Attack	π	η	aux	**Bound**
GRR UNI	$U[m]$	0	–	Corollary 2, Corollary 1
GRR TRUE	π	0	–	Theorem 7, Theorem 8
GRR CORR	π	0	v_{t-1}	Theorem 9
EM UNI	$U[m]$	0 $\{0.1, 0.3, 0.5, 0.7\} \cdot \Delta_G$	–	Corollary 1 Theorem 6
EM CORR	π	0 $\{0.1, 0.3, 0.5, 0.7\} \cdot \Delta_G$	v_{t-1}	Theorem 9

Table 5.4.: Theoretical bounds for each attack setting.

For the theoretical bounds, we have to compute the baseline errors. The average baseline error κ_π is computed according to Equation (4.1). In contrast, the upper baseline error $\kappa^+_{\pi,\eta}$ depends on η and does not have a closed-form solution. We compute it for the Porto and Beijing graphs respectively by iterating over all nodes v of the graph and, for each v, evaluating the probability that a randomly chosen node $Z \sim \pi$ lies within distance η of v,

$$\Pr_{Z\sim\pi}[\ell(v, Z) \leq \eta] = \sum_{w\in\mathcal{V}} \pi(w) \cdot \mathbf{1}\{\ell(v, w) \leq \eta\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. After computing this probability for all nodes, we take the maximum over v to obtain $\kappa^+_{\pi,\eta}$,

$$\kappa^+_{\pi,\eta} = \max_{v\in\mathcal{V}} \Pr_{Z\sim\pi}[\ell(v, Z) \leq \eta],$$

which implements Equation (3.4).

Finally, for the bound in Theorem 6, we have to compute the TV of the GRR and EM mechanisms. For GRR, we have a closed-form solution (cf. Example 1) but for the EM, no closed-form solution exists, so we compute the TV numerically. For each value of ε, we consider the output distributions of the EM for all possible input nodes and identify the pair of nodes whose output distributions are maximally distinguishable. The TV is then obtained as half of the maximum ℓ_1-distance between these distributions [27].

Markov model: For the CORR attack setting, we require a Markov model as background knowledge for the adversary. In order to train a model that learns the global data distribution but not any participant's data we follow the state-of-the-art evaluation approach for membership inference attacks [44], i.e., we split our datasets in $10,000$ trajectories (test dataset) to simulate the target users and use the rest of the dataset (shadow model data) to train the Markov transition model.

6. Results on LDP

In this chapter, we present the experimental results of the attacks conducted under the LDP setting.

6.1. Results for GRR

We first present the results for the attacks on the GRR mechanism. In this setting, we only evaluate perfect reconstruction.

6.1.1. Uniform Prior (UNI)

For the UNI attack, we can compute ReRo and U-ReRo in closed form. Since π is the uniform distribution over the set of m possible locations, an oblivious adversary predicting a location at random location out of the m available nodes and will be correct with probability $\frac{1}{m}$.

The adversary directly outputs the location perturbed with GRR as their prediction and will thus be correct with probability p. Therefore,

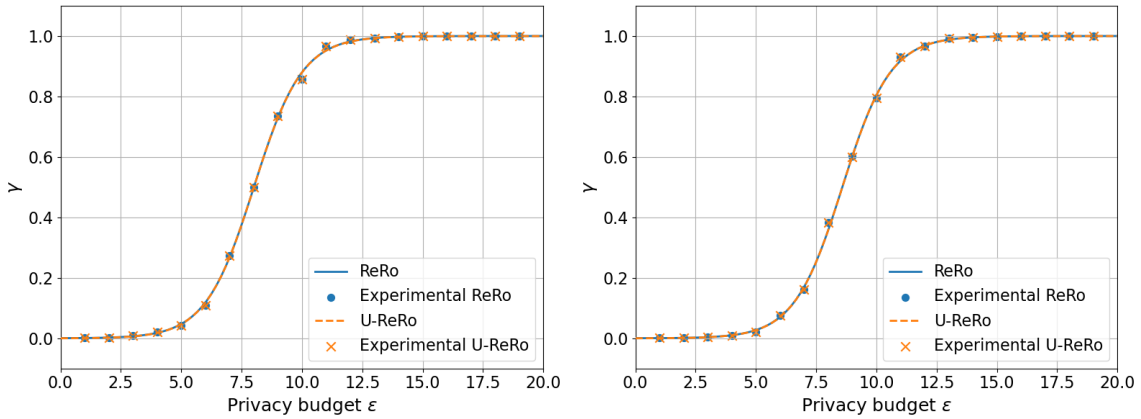
$$\Pr_{\substack{Z\sim U[m]\\ \theta\sim\mathcal{M}(D_Z)}}[\ell(Z,A(\theta))=0] = \Pr_{Z\sim U[m]}[\mathcal{M}_{GRR}(Z)=Z] = p.$$

The U-ReRo correction term is computed as:

$$\mathbb{E}_{Z_0\sim U[m]}\left[\Pr_{Z\sim U[m]}(\mathcal{M}_{GRR}(Z_0)=Z)\right] =$$
$$\mathbb{E}_{Z_0\sim U[m]}\left[\frac{p}{m}+\frac{m-1}{m}q\right] = \frac{1}{m}\left(p+\frac{m-1}{e^{\epsilon}+m-1}\right).$$

In Figure 6.1, we present both the theoretical predictions (derived from the formulas above) and the empirical results for ReRo and U-ReRo. The results are shown for both the Porto and Beijing datasets, which yield similar outcomes. The empirical values closely match the theoretical curves, indicating an accurate approximation of the underlying random variables.

Notably, ReRo and U-ReRo are nearly identical. This is because the adversary lacks access to informative knowledge: whenever the GRR output differs from the target, the adversary is effectively forced to guess at random. Since both graphs are large–$m = 3052$ for Porto and $m = 5356$ for Beijing–the correction term $\frac{1}{m}$ is negligible.



(a) Porto dataset.

(b) Beijing dataset.

Figure 6.1.: Comparing ReRo and U-ReRo based on both theoretical and experimental evaluations for the GRR UNI attack. Theoretical expectations are plotted as lines and experimental results as crosses.

When the adversary has to choose only between $m = 2$ nodes, the probability of a correct guess increases, even when the prediction does not correspond to the actual target. Consequently, the correction term becomes more significant. In Figure 6.2, we plot the results and observe that the correction term now equals 0.5. This confirms that, in the absence of informative background knowledge or *aux*, the correction term reflects only the baseline probability of random guessing. Intuitively, this aligns with the idea that any successful attack under a uniform prior must result from the information leaked from the DP mechanism $\mathcal{M}$.

We plot the theoretical bounds for this setting (cf. Corollary 1 and Corollary 2) with our experimental results in Figure 6.3, for both the Porto and Beijing datasets. In both cases, we observe a perfect alignment between the empirical results and the theoretical prediction. This precise match provides strong evidence that the bound we derived for the U-ReRo metric accurately captures the privacy leakage in this setting. Specifically, it confirms that the bound does not *overestimate* the privacy risk: the observed leakage does not fall significantly below the predicted values, meaning we are not being overly pessimistic in our theoretical estimation.

More importantly, we can also rule out the possibility of *underestimating* the privacy leakage. This is supported by the findings of Arcolezi et al., who show that the UNI attack strategy is optimal against GRR under a uniform data distribution. That is, among all possible inference strategies, the GRR UNI attack achieves the highest possible privacy leakage. Consequently, there cannot exist any tighter bound. We conclude that our theoretical bound precisely characterizes the worst-case privacy leakage without either over- or underestimating it.

Finally, we note that for the GRR mechanism, Corollary 1 and Corollary 2 yield equivalent results. In general, the evaluation of Corollary 2 requires knowledge of the specific DP mechanism to compute its f-function whereas Corollary 1 is applicable even when this information is not given.

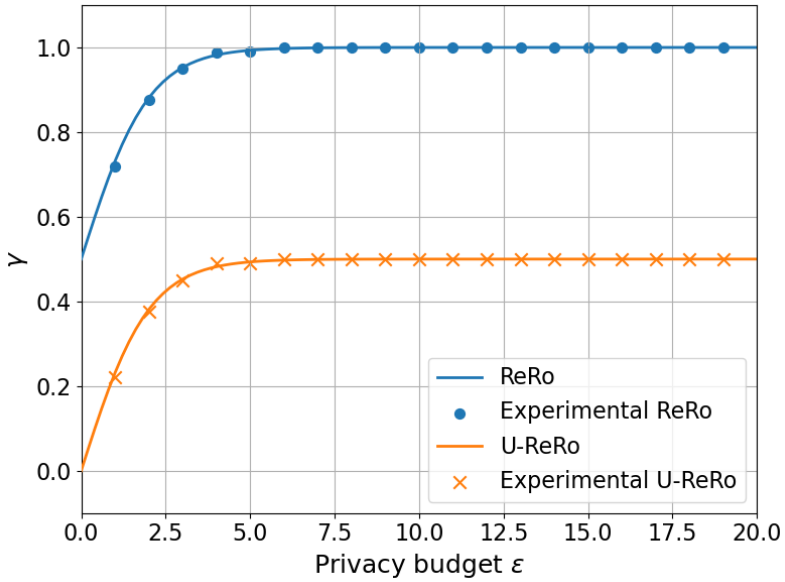
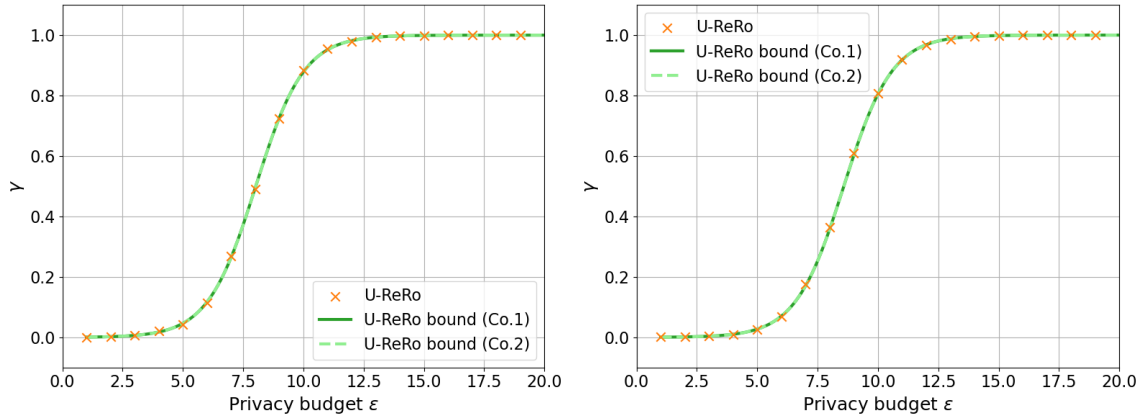


Figure 6.2.: Comparing ReRo and U-ReRo for GRR UNI attack when the adversary has to choose only from $m = 2$ nodes.



(a) Porto dataset.

(b) Beijing dataset.

Figure 6.3.: Comparing the experimental results for GRR UNI attack with the theoretical bounds.

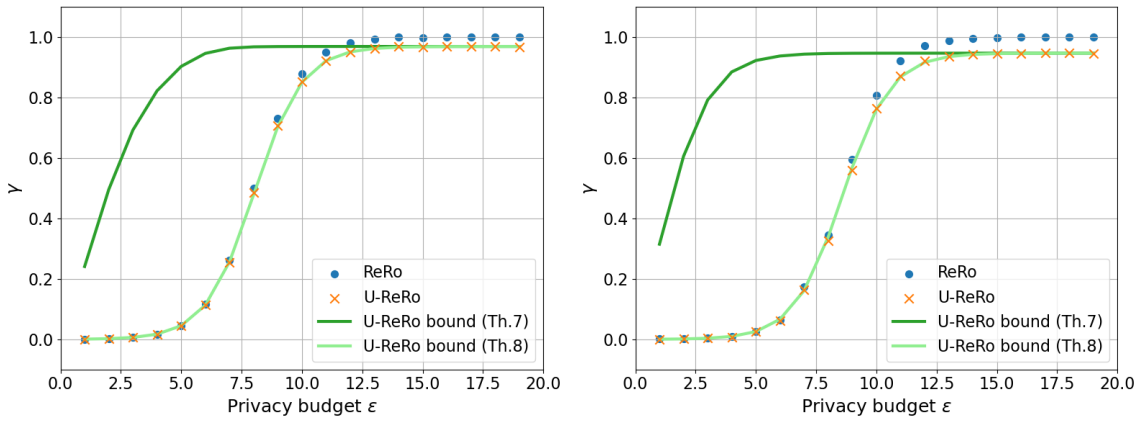
Next, we evaluate the same attack strategy given the true data distribution π, rather than $U[m]$.

6.1.2. True Data Distribution Attack (TRUE)

The results of Attack 1 for a non-uniform data distribution are plotted in Figure 6.4 for both datasets. For low values of the privacy budget, ReRo and U-ReRo coincide as in the UNI setting.

However, as ε grows larger, we observe an emerging gap. For these values, the probability p of telling the truth approaches 1 and so does ReRo where the report and the target always match. For the correction term the location for the report and the target are sampled independently *but according to the data distribution* π, which increases the overall chance that they are the same, compared to when we sample from a uniform distribution. For low values of ε, the report is likely a randomly chosen location other than the true location, so the correction term is close to $\frac{1}{m}$.

More importantly, our bound from Theorem 8 aligns exactly with the observed leakage. In contrast, the general bound—used when the mechanism is unknown—only approximates the true leakage for $\varepsilon \leq 11$. This is expected, since Theorem 7 does not rely on the specific f-function and TV for GRR but rather on worst-case bounds.



(a) Porto dataset.

(b) Beijing dataset.

Figure 6.4.: Comparing the experimental results of GRR TRUE attack with the theoretical bounds.

6.1.3. Prior Recovery Attack (PRIOR)

For the PRIOR attack, we can compute ReRo and U-ReRo in closed-form. As the adversary disregards any output from the GRR mechanism and always outputs the most popular node, ReRo corresponds to the probability with which the most popular node occurs:

$$\Pr_{Z\sim\pi,\theta\sim\mathcal{M}_{GRR}(Z)}[l_G(Z, A(\theta)) = 0] = \max_{v\in V(G)} \pi(v).$$

The computation of the correction term follows is the same since the attack strategy is oblivious to any mechanism output:

$$\mathbb{E}_{Z_0\sim\pi}\left[\Pr_{Z\sim\pi}(Z = w)\right] = \pi(w) = \max_{v\in V(G)} \pi(v).$$

Since the PRIOR attack is oblivious to the output of the mechanism $\mathcal{M}$, the adversary operates effectively in a setting where $\varepsilon = 0$. In this case, the bound evaluates to 0. This indicates that, under such oblivious conditions, U-ReRo is provably minimal.

The experimental results from Table 6.1 show that ReRo corresponds precisely to the probability of the most frequent node in each graph, which is 0.1 for the Porto and 0.15 for the Beijing dataset. The attack strategy is independent of the mechanism's output; therefore, its success rate remains constant regardless of the privacy budget parameter.

More importantly, we observe that U-ReRo is consistently zero. The attack always produces the same prediction without adapting to any information revealed by the mechanism. Thus, the correction term corresponds exactly to ReRo.

$\max_{v' \in V(G)} \pi(v')$	**ReRo**	**U-ReRo**	**U-ReRo bound**
0.11 (Porto)	0.11	0	0
0.15 (Beijing)	0.15	0	0
0.5	0.5	0	0

Table 6.1.: Comparing ReRo and U-ReRo for GRR PRIOR attack when the adversary knows data distribution π and chooses the most common node. Results for settings with varying probability $\max_{v \in V(G)} \pi(v)$ of the most popular location.

While ReRo and U-ReRo already differ in the prior recovery attack on the Porto and Beijing datasets, the distinction becomes clearer under a strongly skewed distribution—for instance, when the most popular node occurs with probability 0.5, ReRo corresponds to 0.5 whereas U-ReRo remains 0 (Table 6.1).

Based on ReRo alone, one might incorrectly infer a substantial privacy risk, since ReRo captures the adversary's success rate in guessing the most frequent node, which in this case simply reflects knowledge of the data distribution rather than information extracted from the mechanism's output. In contrast, U-ReRo remains zero, showing that the mechanism itself does not leak information and highlighting that ReRo can overestimate privacy risk by conflating general knowledge with actual leakage.

Nonetheless, we acknowledge that the PRIOR attack represents an edge case because the adversary does not use the mechanism's output. In practice, it is plausible that an adversary may extract some information from the mechanism's output, which might not necessarily reveal individual user data but could instead uncover global statistics or aggregate patterns. Therefore, for the next analysis, we are particularly interested in exploring a scenario where the adversary learns insights from the mechanism that go beyond static knowledge.

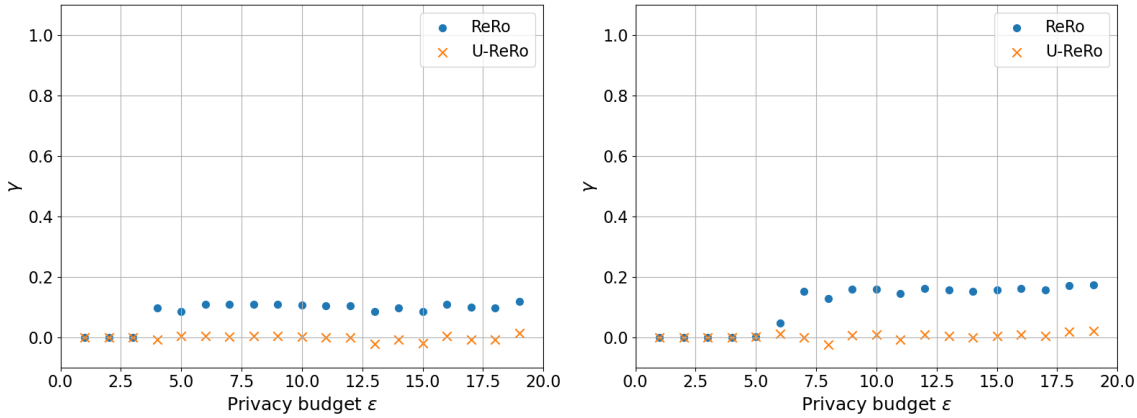
6.1.4. Estimation-based Attack (EST)

We present the results of the attack for both the Porto and Beijing datasets in Figure 6.5. For small values of the privacy parameter ε, the attack performs worse than the PRIOR attack. The adversary relies on a noisy estimate of the data distribution rather than the true distribution when inferring the predicted node. As a result, their predictions are less accurate for low privacy budgets.

As the privacy budget ε increases, we observe that the ReRo metric converges to the probability of the most common node in the respective graph. This behavior mirrors

the outcome observed in the PRIOR setting. The noise added by the mechanism to the collected reported nodes decreases with larger ε, allowing the adversary's estimate of the data distribution to become increasingly accurate and closer to the true distribution.

As in the PRIOR setting, U-ReRo remains consistently close to zero across all values of ε, and it is always lower than ReRo. This finding supports our expectation that although the adversary can learn to predict the most common node by leveraging global statistics learned from the mechanism's output, this does not constitute a privacy violation tied to individual user data. In other words, the adversary's success stems from learning aggregate properties of the data rather than exploiting information specific to any individual user. Consequently, the correction term accounted for in U-ReRo is substantial, leading to an overall U-ReRo value that is effectively zero.



(a) Porto dataset.

(b) Beijing dataset .

Figure 6.5.: Comparing ReRo and U-ReRo for GRR EST attack where the adversary learns an approximate data distribution $\widetilde{\pi}$ from the GRR output and outputs the most popular node based on the estimation.

6.1.5. Correlation-based Attack (CORR)

The CORR attack relies on the parameter τ to control the plausibility filtering using the Markov model. We observe that the attacks are most effective when $\tau = 0$, and their performance deteriorates noticeably even with a small increase, such as $\tau = 0.1$. This suggests that the most effective use of the Markov model is as a binary plausibility filter: checking whether the transition from the previous location to the predicted location, as output by the GRR mechanism, is plausible at all. Based on this insight, we only report results for $\tau \in \{0, 0.1\}$. All parameters for the attack are described in Table 6.2.

In Figure 6.6, we present the results for $\tau = 0$ for both the Porto and Beijing datasets. Three important observations arise from these results:

- We observe a moderate gap between Aux-Aware-ReRo and -U-ReRo. This gap can be attributed to the correction term, which reflects the adversary's ability to fall back on the Markov model when the reported location is not a plausible transition from

Parameter	Meaning	Values
ε	Privacy budget	$\{1, ..., 20\}$
η	Reconstruction threshold	$\{0\}$
I	Number of samples for Monte Carlo estimate	1,000
J	Number of mechanism runs per sample	5
τ	Probability threshold	$\{0, 0.1\}$

Table 6.2.: Parameter settings for CORR attack.

the target user's last known location. Specifically, when an adversary receives a GRR-obfuscated location report from a different user, it is unlikely that this location aligns with the Markov model transition probabilities of the target user. As a result, the Markov model plausibility check fails, and the adversary resorts to sampling a prediction from the model directly. This fallback strategy performs better than uniform random guessing (as observed in the GRR UNI baseline), resulting in a non-trivial success rate and thus explaining the moderate performance gap.

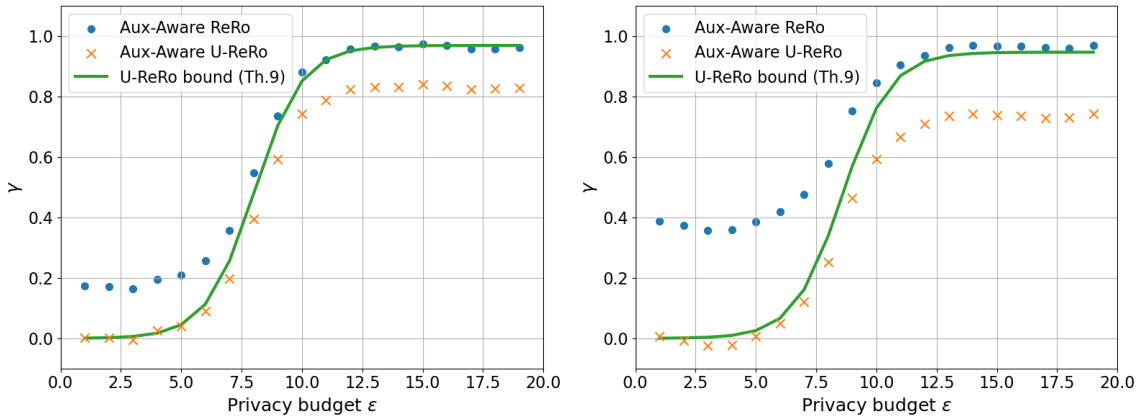
- We also observe that Aux-Aware-U-ReRo is not zero for larger values of the privacy budget ε. This is in contrast to the PRIOR and EST attack settings, where U-ReRo remained consistently around zero, indicating negligible individual leakage. With the GRR mechanism, as ε increases, the probability p of reporting the true location approaches 1. Consequently, users are more likely to reveal their actual locations. When this occurs, the adversary can rely on the reported location, which now passes the Markov model plausibility check, to make an accurate inference. This leads to actual individual privacy leakage that was absent in the previously considered attack settings.

- The theoretical bound from Theorem 9 is tight for small values of $\varepsilon \leq 5$. For larger values ($\varepsilon \geq 10$), however, a consistent gap appears. Because this bound is tailored to the GRR mechanism and its TV, it is unlikely to overestimate privacy leakage; rather, the gap suggests that the attack strategy is suboptimal–particularly in contrast to the performance of the attack in the TRUE setting, which also involved a non-uniform data distribution.

We plot the results for $\tau = 0.1$ in Figure 6.7. We observe worse attack performance for this threshold value but overall the same trends as discussed above.

Parameter	Porto dataset	Beijing dataset
κ_π	0.03	0.05
$\kappa^+_{\pi,\eta}$	0.10	0.15

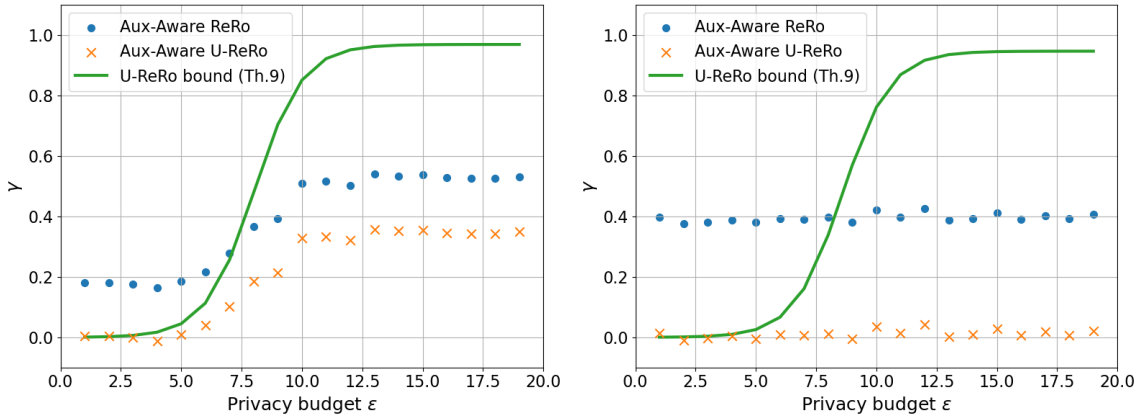
Table 6.3.: Average and upper baseline error of data distribution π for Porto and Beijing datasets ($\eta = 0$).



(a) Porto dataset.

(b) Beijing dataset .

Figure 6.6.: Comparing Aux-Aware ReRo and -U-ReRo for GRR CORR attack with threshold $\tau = 0$ for transition between previous location and perturbed location.



(a) Porto dataset.

(b) Beijing dataset.

Figure 6.7.: Comparing Aux-Aware ReRo and -U-ReRo for GRR CORR attack with threshold $\tau = 0.1$ for transition between previous location and perturbed location.

6.2. Results for EM

For the attacks on the EM, we evaluate both perfect and partial reconstruction.

6.2.1. Uniform Prior (UNI)

We can compute ReRo and U-ReRo in closed-form for the UNI attack. With the EM, for any node $v \in V(G)$, the perturbed node $\widetilde{v} \in V(G)$ is sampled according to

$$\widetilde{v} \sim \exp\left(\frac{-\varepsilon l_G(v, \widetilde{v})}{2\Delta_G}\right),$$

where Δ_G is the diameter of the graph.

ReRo is then computed as:

$$\Pr_{Z\sim U[m]}[\mathcal{M}_{Exp}(Z) = Z] = \sum_{v\in V(G)} \frac{1}{|V(G)|} \cdot \frac{\exp\left(\frac{-\varepsilon l_G(v,w)}{2\Delta_G}\right)}{\sum_{w'\in V(G)} \exp\left(\frac{-\varepsilon l_G(v,w')}{2\Delta_G}\right)}$$
$$\overset{l_G(v,w)=0}{=} \sum_{v\in V(G)} \frac{1}{|V(G)|} \cdot \frac{1}{\sum_{w'\in V(G)} \exp\left(\frac{-\varepsilon l_G(v,w')}{2\Delta_G}\right)}.$$

The U-ReRo correction term is equal to:

$$\mathbb{E}_{Z_0\sim U[m]}\left[\Pr_{Z\sim U[m]}[\mathcal{M}_{Exp}(Z_0) = Z]\right]$$
$$= \frac{1}{|V(G)|} \sum_{v\in V(G)} \Pr_{Z\sim U[m]}[\mathcal{M}_{Exp}(v) = Z]$$
$$= \frac{1}{|V(G)|^2} \sum_{v\in V(G)} \sum_{w\in V(G)} \frac{\exp\left(\frac{-\varepsilon l_G(v,w)}{2\Delta_G}\right)}{\sum_{w'\in V} \exp\left(\frac{-\varepsilon l_G(v,w')}{2\Delta_G}\right)}.$$

We list the experimental parameters in Table 6.5. The values for the baseline errors for both datasets can be found in Table 6.4.

Porto			**Beijing**		
η	$\kappa^+_{\pi,\eta}$	κ_π	η	$\kappa^+_{\pi,\eta}$	κ_π
0	0.10	0.03	0	0.15	0.05
9	0.29	0.03	9	0.34	0.05
26	0.84	0.03	28	0.71	0.05
44	1	0.03	46	0.99	0.05
70	1	0.03	74	1	0.05

Table 6.4.: Average and upper baseline error of data distribution π for Porto and Beijing datasets for varying η.

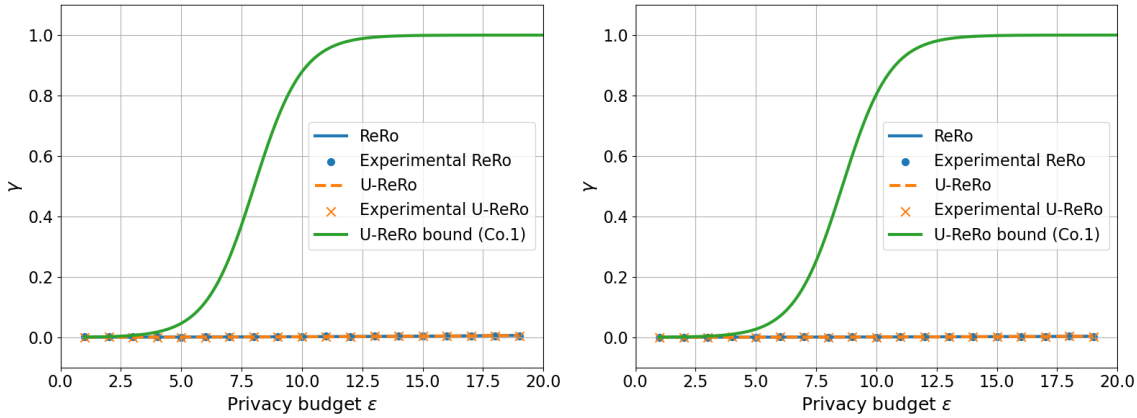
Parameter	**Meaning**	**Values**
ε	Privacy budget	$\{1, ..., 20\}$
η	Reconstruction threshold	Porto: $\{0, 9, 26, 44, 70\}$ Beijing: $\{0, 9, 28, 46, 74\}$
I	Number of samples for Monte Carlo estimate	1,000
J	Number of mechanism repetitions per sample	5
τ	Probability threshold (CORR attack)	0

Table 6.5.: Parameter Settings for UNI and CORR Attacks.

The results in Figure 6.8 show that ReRo and U-ReRo are almost equal under uniform prior, as in the GRR UNI attack. However, in contrast to these previous results, we find

that the EM UNI attack achieves negligible performance across the entire range of privacy budgets ε considered in our experiments. This results from the scale of the noise introduced in the EM based on the sensitivity Δ_G, the graph diameter. For the Porto dataset, this results in a sensitivity of $\Delta_G = 88$, while for the Beijing dataset, $\Delta_G = 92$, which leads to substantial noise. Consequently, the likelihood of the adversary correctly guessing the exact true location by using the EM's noisy output as prediction becomes very low.

In order to verify that our attack strategy works correctly, we carried out the attack on a smaller synthetic graph with $m = 50$ nodes and a diameter $\Delta_G = 11$ which results overall in less noise added to the location reports. In Figure 6.9, we can see that the attack indeed performs better and we see ReRo and U-ReRo rise moderately for lower values of ε. As expected, the U-ReRo correction term results in $\frac{1}{m} = \frac{1}{50}$, discounting only for random guessing. We also note that our experimental results closely track the closed-form solution for the EM UNI attack as derived above.



(a) Porto dataset ($\Delta_G = 88$). (b) Beijing dataset ($\Delta_G = 92$).

Figure 6.8.: Comparing ReRo and U-ReRo for EM UNI attack for $\eta = 0$.

The results for partial reconstruction and varying reconstruction thresholds η are shown in Figure 6.10 for the Porto dataset and in Figure 6.11 for the Beijing dataset. We observe that the difference between ReRo and U-ReRo increases as η grows. For ReRo, the adversary's performance improves steadily. This can be explained by the fact that the EM is more likely to output nodes that are closer to the input in terms of path distance. Since the adversary simply outputs the perturbed value as their prediction, its success rate increases as η grows.

In contrast to ReRo, U-ReRo decreases as η increases due to its growing correction term. Even though the adversary's prediction under U-ReRo is based on a perturbed report for a different target, the attack can still succeed–simply because more nodes qualify as acceptable reconstructions when η is large. As a result, even random predictions can fall within the allowed radius, boosting success rates.

However, U-ReRo is not zero for intermediate values of η, which signals that some information still leaks from the mechanism's output. The EM tends to output nodes closer to the true location, based on the scoring function. When the prediction is the perturbed report of the actual target (as in ReRo), it is more likely to fall within η hops, increasing

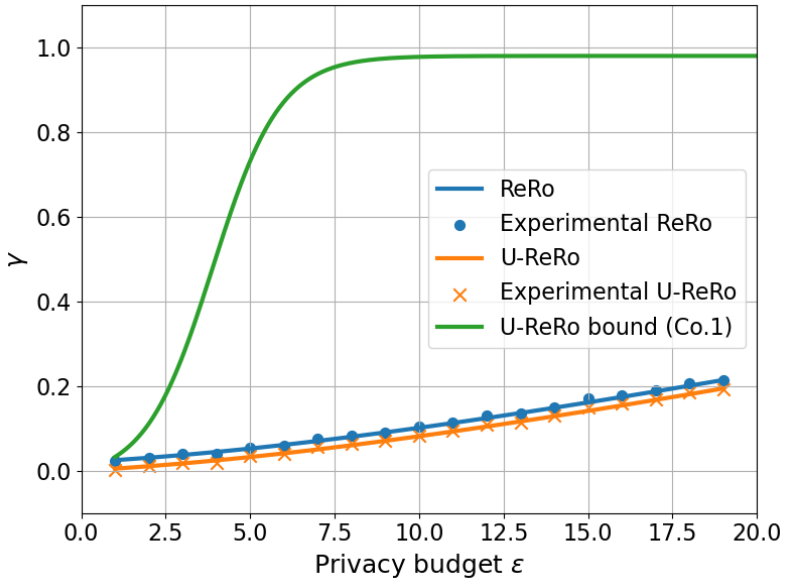


Figure 6.9.: EM UNI attack on synthetic graph with $m = 50$ nodes and a diameter of $\Delta_G = 11$ for $\eta = 0$.

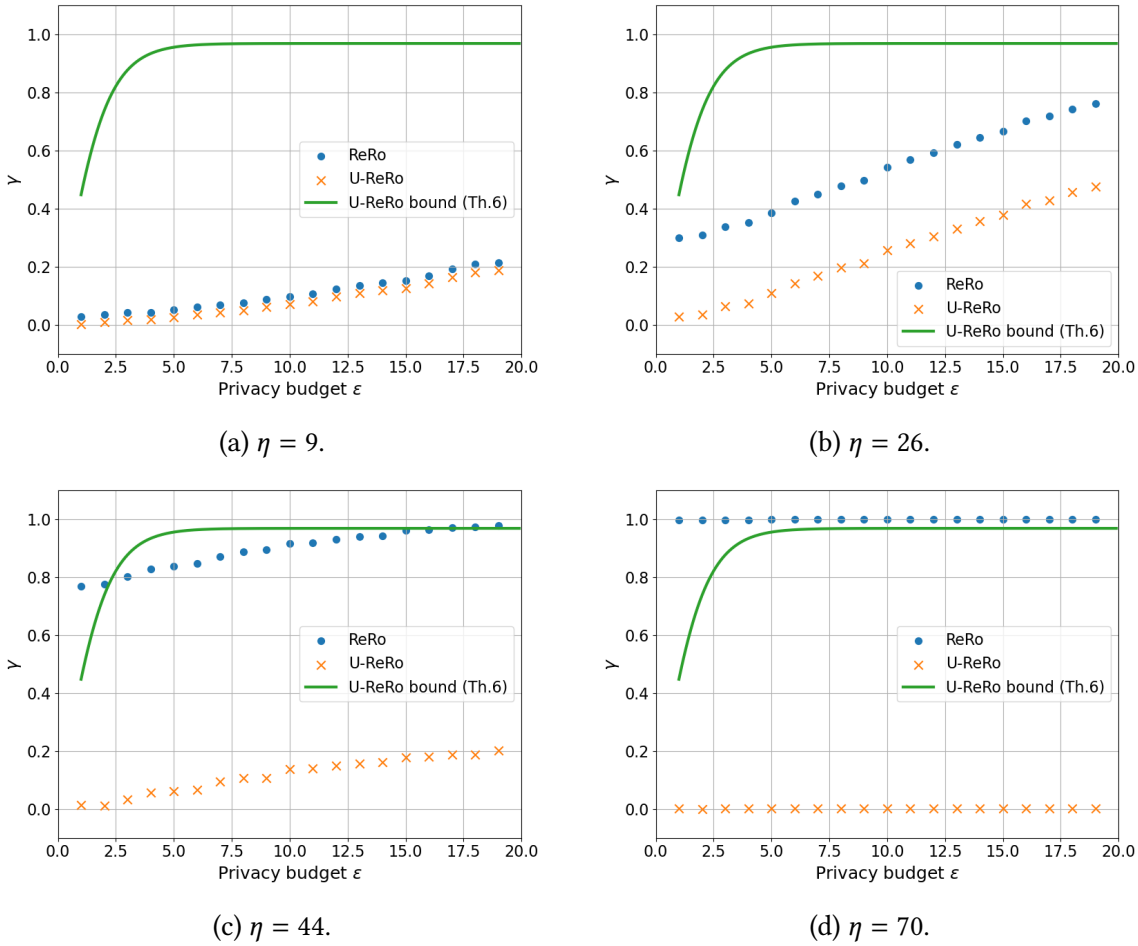
the success rate. Not all of this success is accounted for by the correction term. This means users are exposed to some risk: even a noisy report can reveal coarse-grained information about their true location. As ε increases, this leakage becomes more pronounced, especially for intermediate η values–demonstrating that the adversary learns non-trivial information about the target.

When η becomes too large, the task becomes trivial–nearly any prediction counts as correct. Then, U-ReRo drops to zero, indicating that there is no meaningful attack success. ReRo, in contrast, always assigns a success of 1, which overestimates the adversary's capabilities and fails to account for the triviality of the task.

6.2.2. Correlation-based Attack (CORR)

The results for perfect reconstruction and probability threshold $\tau = 0$ can be found in Figure 6.12. We observe a moderate difference between Aux-Aware ReRo and -U-ReRo, as we did for the GRR CORR attack. However, the overall success probability of the adversary is much lower which results from the high degree of perturbation applied by the EM mechanism due to the large global sensitivities of the graphs. Even for large ε, the adversary's success chances do not improve. Furthermore, the Aux-Aware U-ReRo values are consistently around 0, indicating that any success is due to the Markov model and the target-specific auxiliary knowledge *aux*, rather than the output of the mechanism.

In the partial reconstruction setting, we observe a trend consistent with the EM UNI attack: as the parameter η increases, so does the gap between Aux-Aware ReRo and -U-ReRo. This pattern is illustrated in Figure 6.13 for the Porto dataset and in Figure 6.14 for the Beijing dataset. Notably, the Aux-Aware U-ReRo values remain close to zero



(a) $\eta = 9$.

(b) $\eta = 26$.

(c) $\eta = 44$.

(d) $\eta = 70$.

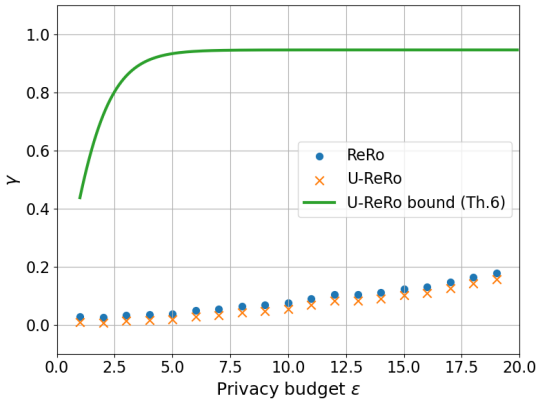
Figure 6.10.: EM UNI on Porto dataset for varying reconstruction thresholds η.

across different values of η, indicating that the adversary's success is not attributable to information learned from the mechanism's output.

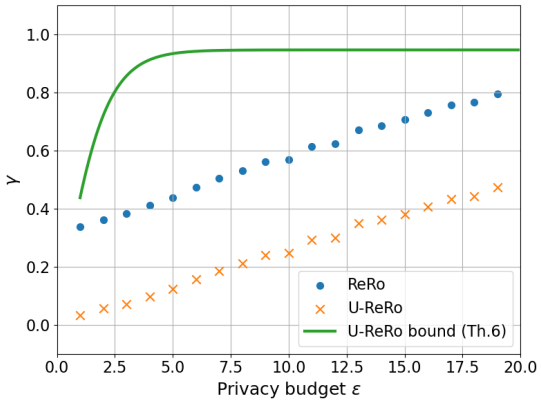
Our attack strategy outputs the perturbed node as the prediction, which increases the success probability when the predicted and true nodes coincide. However, the significant correction term reflected in Aux-Aware U-ReRo underscores the strength of the adversary's knowledge of *aux* and the Markov model. Even when the perturbed node used for prediction and the actual target are sampled independently, the adversary retains a high probability of successful reconstruction. This becomes especially evident for larger values of η (e.g., $\eta = 44$ and $\eta = 70$), where the reconstruction task becomes trivial. In these cases, the adversary achieves near-perfect success without requiring any additional information from the mechanism's output.

6.3. Intermediate Conclusions

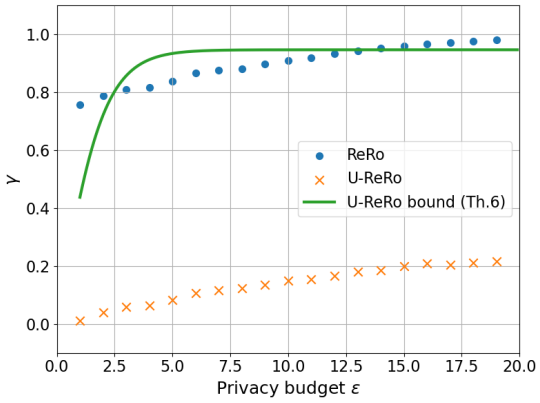
Our experiments on LDP mechanisms confirm that ReRo fails to distinguish between individual privacy leakage–arising from participation in the dataset–and adversarial success



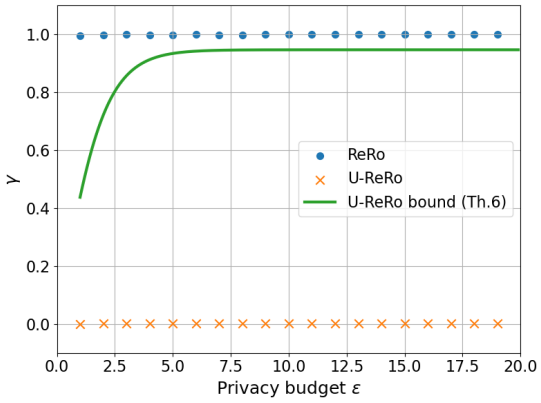
(a) $\eta = 9$.



(b) $\eta = 28$.

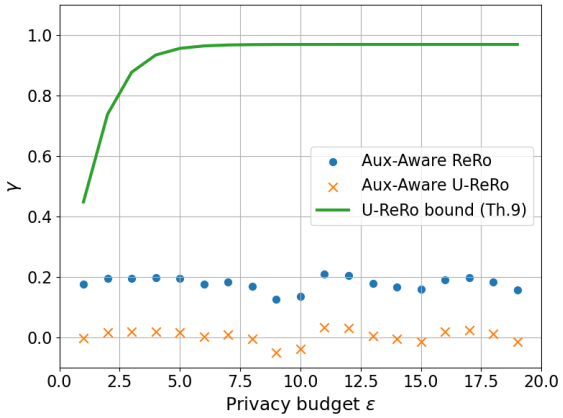


(c) $\eta = 46$.

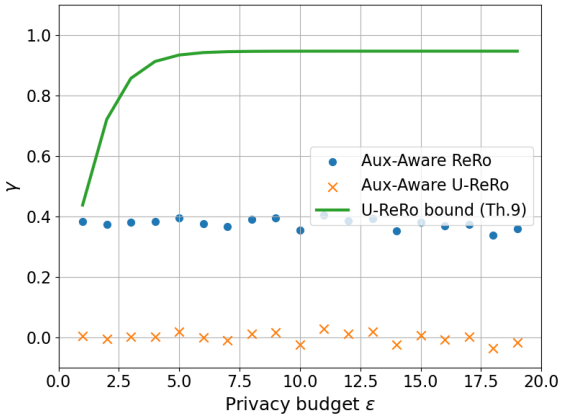


(d) $\eta = 74$.

Figure 6.11.: EM UNI on Beijing dataset for varying reconstruction thresholds η.

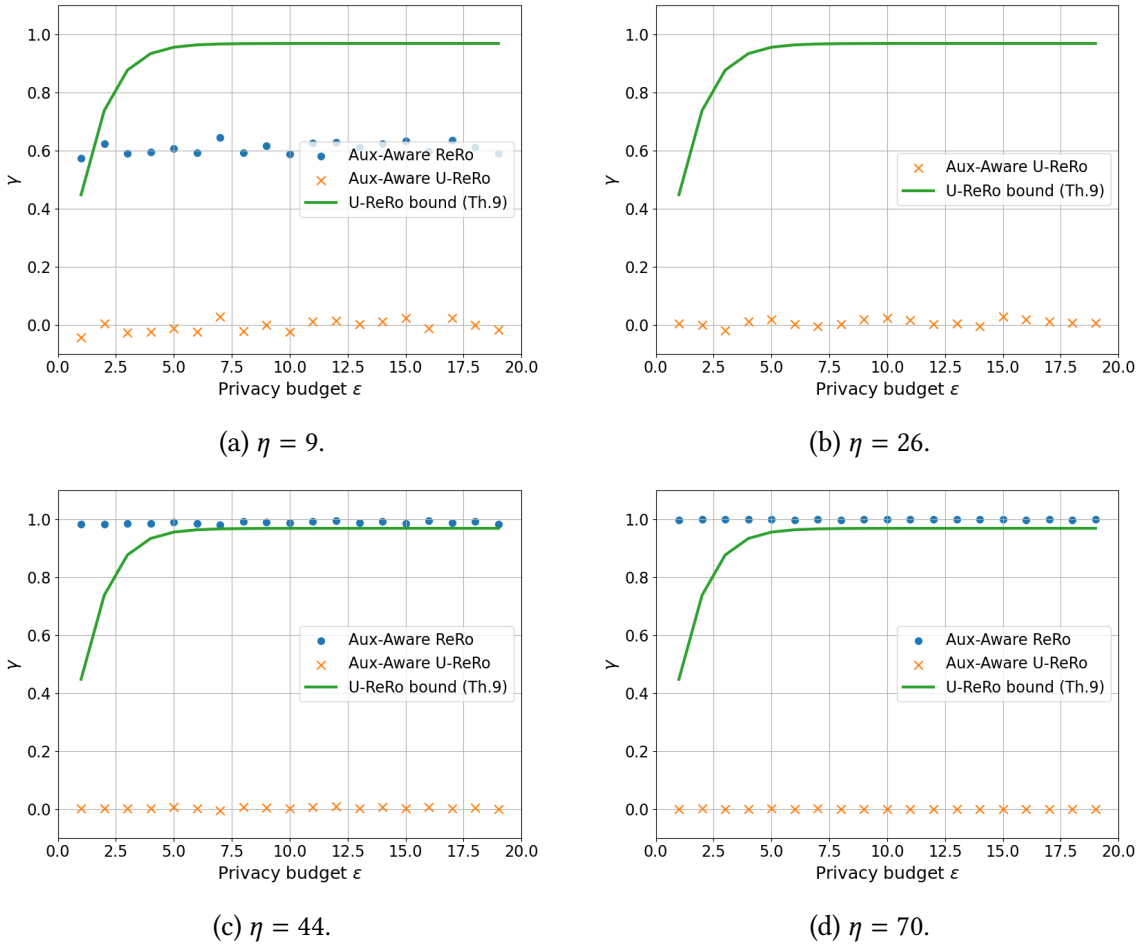


(a) Porto dataset ($\Delta G = 88$).



(b) Beijing dataset ($\Delta G = 92$).

Figure 6.12.: Comparing Aux-Aware ReRo and -U-ReRo for EM CORR attack for $\eta = 0$ (threshold value $\tau = 0$ for Markov model plausibility check).



(a) $\eta = 9$.

(b) $\eta = 26$.

(c) $\eta = 44$.

(d) $\eta = 70$.

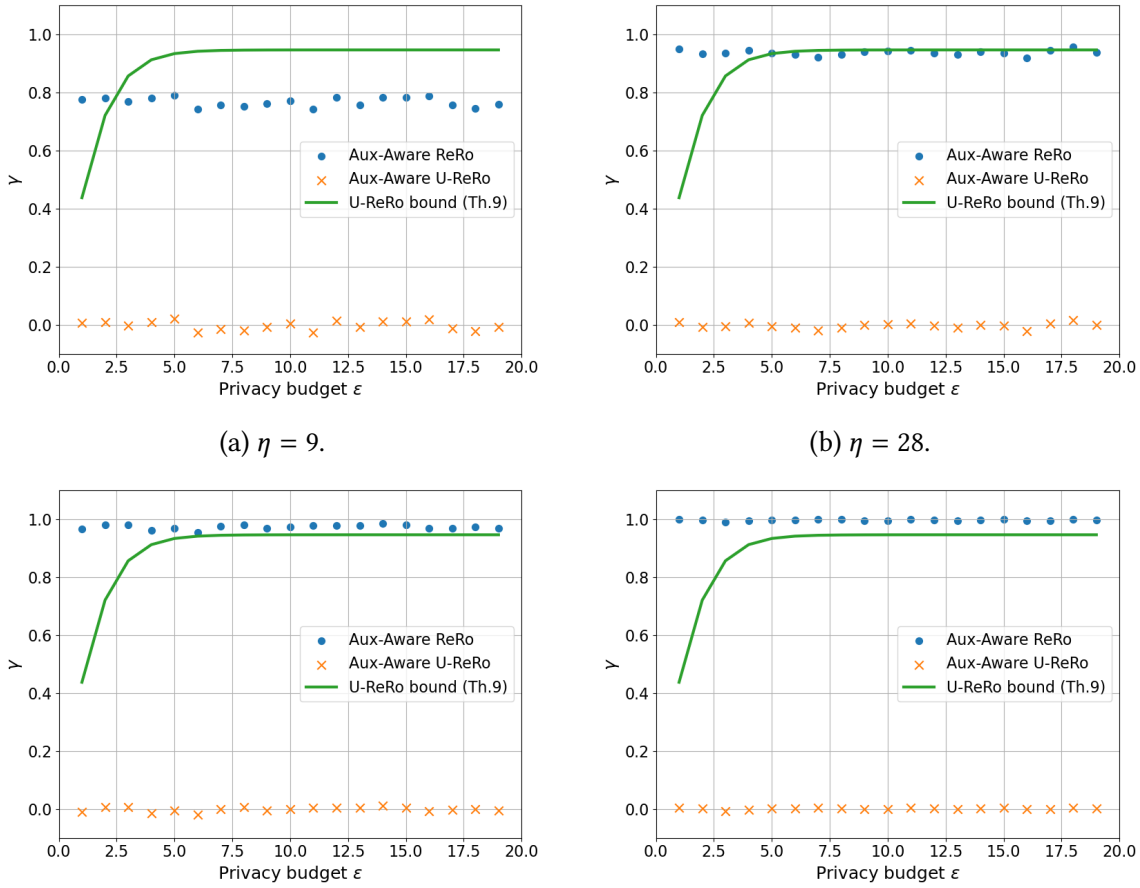
Figure 6.13.: EM CORR on Porto dataset for varying reconstruction thresholds η (threshold $\tau = 0$ for Markov model plausibility check).

based on background knowledge, *aux* or imputatiom. This inflates perceived privacy risks, potentially leading to overly conservative privacy budgets and reduced data utility without corresponding privacy gains.

We saw that U-ReRo successfully isolated privacy leakage resulting from participation in the data collection and discounted for attack success from other sources. We also showed that under uniform priors over large domains, U-ReRo closely tracks ReRo, providing evidence that U-ReRo does not underestimate risk in realistic scenarios.

We furthermore confirmed that our definition of Aux-Aware U-ReRo allows to capture privacy leakage in settings where the adversary has access to *aux*.

For both the GRR UNI and GRR TRUE attacks, we demonstrated that our theoretical bounds closely track the actual attack performance. This is particularly insightful in the case of GRR UNI, as the attack is known to be nearly optimal in this setting. Consequently, our bound provides a tight estimate of the worst-case adversarial success, showcasing its use in guiding privacy budget selection. For the GRR CORR attack, we showed that our bound on Aux-Aware U-ReRo also provided a good estimate of the actual privacy leakage, especially for low values of the privacy budget.



(a) $\eta = 9$.

(b) $\eta = 28$.

(c) $\eta = 46$.

(d) $\eta = 74$.

Figure 6.14.: EM CORR on Beijing dataset for varying reconstruction thresholds η (threshold $\tau = 0$ for Markov model plausibility check).

In experiments with the EM, we observed that the bounds were generally looser. This does not suggest that the bound itself is inadequate; rather, it reflects that the attack strategy optimal for the GRR mechanism does not directly translate to the EM. More importantly, we showed that when $\eta > 0$, (Aux-Aware) ReRo once again overestimates privacy leakage, as it fails to account for the fact that more predictions fall into the accepted error range. In contrast, the (Aux-Aware) U-ReRo correction term increases with η and accounts for the fact that the attack task becomes easier when more reconstructions are considered acceptable.

Part II.

Analyzing U-ReRo in Machine Learning

7. Introduction

We shift focus from the local to the central model of DP, where data is first collected by a trusted curator and obfuscated after the collection. This model is applied by some of the most important practical deployments of DP, including those in official statistics [3] and private ML [1]. Central DP is especially relevant to the training of ML models on sensitive data, as it enables the release of trained models while offering formal privacy guarantees about what they can disclose about their training data [1, 78].

Yet, despite its strong theoretical foundations, differentially private ML has proven vulnerable to a range of sophisticated privacy attacks that can compromise individual-level information in trained models [18, 45, 32, 10]. These attacks have evolved beyond early membership and attribute inference attacks [84, 97, 48] to full data reconstruction [10, 42].

ReRo was first defined in the context of attacks on private learning [10]. In these attacks, the adversary has access to an ML model trained with a differentially private training algorithm, most commonly DP-SGD [1]. The adversary's goal is to recover an entire record from the model's training dataset.

As U-ReRo builds directly on ReRo, private learning presents an important test case for our analysis. Notably, recent work by Hayes et al. introduces a novel white-box attack that closely approaches the theoretical ReRo bound from Theorem 3. In this attack, the adversary has full access to all intermediate gradients of a neural network image classifier and leverages this information to reconstruct images from the training dataset [42]. The demonstrated strength of this attack makes it an ideal benchmark for evaluating U-ReRo and assessing the tightness of our own theoretical bounds in the context of private learning and approximate DP.

We also discuss the role of adversarial capabilities in privacy attacks–particularly the contrast between *white-box* and *black-box* attacks. White-box attacks assume full access to the model's internals, such as parameters or gradients. In contrast, black-box attacks rely solely on query access to the target model, without any insight into its internal state [46]. In a critical evaluation of the effectiveness of black-box attacks, Jayaraman et al. show that many black-box attacks fail to provide meaningful reconstructions beyond simple data imputation–that is, inferring a sensitive attribute based on correlations with known attributes [47]. In these cases, the adversary's success does not imply that the model has leaked individual information but rather that it has learned general statistical patterns.

In our evaluation of privacy attacks in LDP, we showed that ReRo can overestimate privacy leakage when adversarial success is driven by population-level trends rather than individual participation in the dataset. U-ReRo addressed this limitation and correctly accounted for it. We aim to extend this approach to the setting of black-box ML attacks. Specifically, we ask whether U-ReRo can reveal similar limitations in ReRo's interpretation of black-box attack success, and whether it supports the conclusions of Jayaraman et al.–namely, that many black-box attacks exploit correlations rather than leaking individual-

level information. By doing so, we provide a more accurate metric for evaluating privacy risks in ML.

7.1. Mechanism Selection: DP-SGD for Classification Models

We employ DP-SGD as our differential privacy mechanism due to its widespread adoption and strong theoretical guarantees in private machine learning [1]. Described as the "workhorse" of differentially private ML [10, 78], DP-SGD has become a standard method for training private neural networks. In this work, we apply DP-SGD to supervised classification tasks, aiming to learn models that predict class labels from input features. We now provide a brief overview of both classification models and DP-SGD.

A classification model $\mathcal{M}_D$ is trained on a dataset $D = \{(z_i, y_i)\}_{i=1}^{|D|}$, where each input $z_i \in \mathcal{Z}$ is associated with a label $y_i \in \mathcal{Y}$. The model corresponds to a function $f_\theta : \mathcal{Z} \to \mathcal{Y}$, parameterized by weights θ, which are learned by minimizing the empirical loss over D using a loss function $\ell(\theta, z)$. We denote by $\hat{y}$ the label predicted by $\mathcal{M}_D$ for a record z, and by $\mathbf{c}$ the associated *confidence vector* over classes $y \in \mathcal{Y}$.

Classification tasks are among the most common ML applications in practice and are widely used in privacy-sensitive domains such as healthcare, finance, and demographics [84, 45, 48]. These settings demand strong guarantees that training data cannot be exposed through a published model. A standard method for privacy-preserving training is *stochastic gradient descent* (SGD) [57], which updates parameters using gradients computed over a batch $B \subseteq D$:

$$\theta_{i+1} \leftarrow \theta_i - r \cdot \frac{1}{|B|} \sum_{z \in B} \nabla_\theta \ell(\theta_i, z),$$

where r is the learning rate determining the step size in each iteration.

DP-SGD extends the standard SGD algorithm by clipping the per-sample gradients and then applying the GM (cf. Section 3.1.1) to the aggregated gradients [1]. Given a sampling probability q, the algorithm forms a batch B by including each example independently with probability q. The per-example gradients are clipped to a fixed norm bound C, and Gaussian noise is added. The update rule becomes [1]:

$$\theta_{i+1} \leftarrow \theta_i - r \cdot \frac{1}{|B|} \sum_{z \in B} \text{clip}_C(\nabla_\theta \ell(\theta_i, z)) + \mathcal{N}(0, C^2\sigma^2\mathbb{I}).$$

The clipping function $\text{clip}_C(g)$ scales a gradient vector g such that its L_2-norm does not exceed C [1]:

$$\text{clip}_C(g) = \frac{g}{\max(1, \|g\|_2/C)}.$$

Abadi et al. show that this procedure achieves (ε, δ)-DP, with

$$\varepsilon \approx \frac{q\sqrt{T \log(1/\delta)}}{\sigma},$$

where T is the number of training steps and $q = \frac{|B|}{|D|}$ the batch sampling ratio [1].

Hayes et al. prove a specific bound for ReRo under DP-SGD. They employ Theorem 3 but instead of comparing the output distributions of $\mathcal{M}_{DP-SGD}$ on databases D, D', they reduce the problem to comparing the T-dimensional Gaussian distribution $Q_{\sigma,q} = \mathcal{N}(0, \sigma^2 I)$ with the Gaussian mixture $P_{T,\sigma,q} = \sum_{w \in \{0,1\}^T} \Pr[\mathrm{Ber}(q,T) = w]\mathcal{N}(w, \sigma^2 I)$ where $\mathrm{Ber}(q,T)$ is a binary vector with each coordinate sampled independently from a Bernoulli distribution with success probability q [62]. This then yields the following bound:

Corollary 3 (ReRo bound for DP-SGD [42]) *$\mathcal{M}_{DP-SGD}$ is (η, γ)-ReRo with*

$$\gamma = 1 - T(P_{T,\sigma,q}, Q_{\sigma,q})(\kappa^{+}_{\pi,\eta}).$$

7.2. Attack Descriptions

We evaluate the state-of-the-art white-box attack [42], which closely approaches the theoretical ReRo bound and serves as a strong test case for evaluating the tightness of our own bounds. As a next step, we implement a data imputation attack that exploits population-level correlations without accessing any private information [47] to evaluate whether ReRo conflates imputation success with privacy leakage and whether U-ReRo correctly distinguishes this. Finally, we reproduce several black-box attacks studied by [47] to investigate whether U-ReRo confirms their finding that such attacks do not succeed beyond imputation. All attacks are described in detail in the following section.

7.2.1. White-box Attack

Hayes et al. present a white-box data reconstruction attack against DP-SGD in which the adversary has access to all intermediate privatized gradients $\{g_1, ..., g_T\}$ released during training. The goal of the attack is to reconstruct a target record $z^* \in D$ from the training dataset by leveraging its influence on these gradients.

The authors assume an *informed adversary* who knows the full training dataset D except for the single unknown target record z^*. This known subset is denoted $D_- = D \setminus \{z^*\}$. Additionally, the adversary is given a discrete candidate set $\{z_1, ..., z_m\}$ of possible values for z^* The task is then to identify the true z^* among the candidates.

The core idea is to simulate the effect that each candidate z_i would have had on the training gradients of the model and compare this to the observed privatized gradients. In DP-SGD, each privatized gradient g_t at training step t is the sum of clipped gradients of individual training examples with added Gaussian noise. The adversary subtracts the known contributions of the records in D_- to isolate the gradient information attributable to z^*:

$$\bar{g}_t = g_t - \sum_{z \in D_-} \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z)).$$

This residual $\bar{g}_t$ approximates the clipped gradient of the unknown z^* plus noise.

For each candidate z_i, the adversary then computes its clipped gradient $\mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))$ and measures how well it aligns with the noisy gradient $\bar{g}_t$ by calculating their inner

product. Summing these inner products over all training steps gives the score

$$s_i = \sum_{t=1}^{T} \left\langle \text{clip}_C(\nabla_\theta \ell(\theta_t, z_i)), \bar{g}_t \right\rangle .$$

The noisy gradient at step t is given by

$$\bar{g}_t = \text{clip}_C(\nabla_\theta \ell(\theta_t, z^*)) + \mathcal{N}(0, C^2\sigma^2 I),$$

where the noise is isotropic Gaussian with covariance scaled by $C^2\sigma^2$. Because the noise is isotropic, when we project it onto the fixed vector $\text{clip}_C(\nabla_\theta \ell(\theta_t, z_i))$, which has norm at most C, the noise term behaves like a one-dimensional Gaussian with variance at most $C^4\sigma^2$.

Thus, the inner product is distributed as

$$\left\langle \text{clip}_C(\nabla_\theta \ell(\theta_t, z_i)), \bar{g}_t \right\rangle \sim \begin{cases} \mathcal{N}(C^2, C^4\sigma^2), & \text{if } z_i = z^* \\ \mathcal{N}(A, C^4\sigma^2), & \text{otherwise} \end{cases}$$

where $A < C^2$ represents the lower expected alignment for incorrect candidates.

By summing these inner products over all T steps, the correct candidate's score s_i tends to be significantly higher than the others, allowing the adversary to identify the true target record [42]. The attack strategy is summarized in Attack 5.

Attack 5 (Prior-aware gradient-based attack [42]) *Let $\{z_1, ..., z_m\}$ be a discrete prior over candidate records from $\mathcal{Z}$, θ_t the model parameters at step $t \in [T]$, and $\{g_1, ..., g_T\}$ the privatized gradients released by DP-SGD. The adversary proceeds as follows:*

- *For each candidate $z_i \in \{z_1, ..., z_m\}$:*
  - *Computes the clipped gradient $\text{clip}_C(\nabla_\theta \ell(\theta_t, z_i))$ for each step t.*
  - *Computes the residual gradient:*

$$\bar{g}_t = g_t - \sum_{z \in D_-} \text{clip}_C(\nabla_\theta \ell(\theta_t, z)).$$

  - *Computes the score:*

$$s_i = \sum_{t=1}^{T} \left\langle \text{clip}_C(\nabla_\theta \ell(\theta_t, z_i)), \bar{g}_t \right\rangle .$$

- *Outputs the candidate z_i with the highest score s_i.*

We evaluate this attack both in its original setting under a uniform data distribution, where each record from the candidate set is equally likely, and under a non-uniform data distribution.

7.2.2. Imputation Attack and Black-box Attacks

For the imputation and black-box attacks, we analyze attribute inference as a form of data reconstruction. Each data point is a pair $z = (\mathbf{x}, s)$, where $\mathbf{x}$ represents the public attributes and s the sensitive attribute to be inferred. The adversary is given access to the $\mathbf{x}$ as *aux* and aims to infer $s \in \mathcal{S}$. This setup corresponds to a classification problem over $|\mathcal{S}|$ possible classes, i.e., $m = |\mathcal{S}|$, where the attacker attempts to perfectly reconstruct the sensitive attribute. In this setting, we distinguish between the classification task of the target model that assigns a label $\hat{y} \in \mathcal{Y}$ for a given record z, and the attack task that is to infer the sensitive attribute s of z while given access to its public attributes $\mathbf{x}$.

We now present the imputation attack and the black-box attacks in detail.

Data imputation originates in the statistics and survey literature, where it was introduced to address the problem of filling in missing values [83, 82], not as an attack but as a method for data recovery. This strategy can be adapted as an attribute inference attack that does not interact with the model trained under DP [47]. Instead, the adversary uses an auxiliary public dataset D_{aux} that approximates the true distribution π of the private training data D. Using D_{aux}, a separate ML model is trained to predict the sensitive attribute from the non-sensitive features. The formal attack procedure is provided in Attack 6.

Attack 6 (Attribute Inference via Data Imputation [47]) *Let $\mathcal{Z}$ denote the data domain, where each record $z = (\mathbf{x}, s) \in \mathcal{Z}$ consists of a sensitive attribute $s \in \mathcal{S}$ and a set of non-sensitive attributes $\mathbf{x} \in \mathcal{X}^r$. An adversary A, with access to a public dataset $D_{aux} \sim \pi$, trains an imputation model $\mathcal{I} : \mathcal{X} \to \mathcal{S}$ to predict the sensitive attribute from the non-sensitive ones. At inference time, the adversary is given a target record z with only $\mathbf{x}$ revealed and outputs the prediction*

$$\widetilde{s} = \arg\max_{s \in \mathcal{S}} \Pr[s \mid \mathbf{x}],$$

where the conditional distribution $\Pr[s \mid \mathbf{x}]$ is estimated by $\mathcal{I}$ using D_{aux}.

In contrast to the data imputation setting, for the black-box ML attacks the adversary queries ML model $\mathcal{M}_D$ to perform their attack.

The first black-box attribute inference attack was proposed by Yeom et al. [97]. In this attack, the adversary exploits the confidence scores of a trained classification model $\mathcal{M}_D$ to infer the sensitive attribute. Specifically, for a target individual with known features $\mathbf{x}$, the adversary constructs candidate records by pairing $\mathbf{x}$ with each possible value $s_i \in \mathcal{S}$ of the sensitive attribute.

For each candidate record $(\mathbf{x}, s_i)$, the adversary queries the target model $\mathcal{M}_D$ and inspects the model's confidence in classifying the input as s_i. The intuition is that the model tends to assign higher confidence to records it has seen during training. Therefore, if a particular candidate value s_i yields a higher confidence, it is more likely to be the true value. This outcome is then weighed by the prior probability of each sensitive attribute value $Pr[s_i]$. The formal attack strategy is given in Attack 7.

Attack 7 (Yeom Attack [97]) *Given the non-sensitive features $\mathbf{x}$ of a target record $z = (\mathbf{x}, s)$, and the prior distribution $\Pr[s_i]$ over the sensitive attribute values $s_i \in \mathcal{S}$, the adversary:*

- *For each candidate $s_i \in \mathcal{S}$, constructs the record $(\mathbf{x}, s_i)$ and evaluates the model's confidence $\mathbf{c}_i \leftarrow \mathcal{M}_D(z_i = (\mathbf{x}, s_i))$.*

- *Outputs*

$$\widetilde{s} = \arg\max_{s_i \in \mathcal{S}} \Pr[s_i] \cdot \mathbf{c}_i.$$

Unlike the Yeom attack, which estimates membership via confidence scores, the Confidence Score-Based Model Inversion Attack (CSMIA) leverages the model's predicted class label $\hat{y}$ [66]. Given access to the non-sensitive features $\mathbf{x}$ of a target z and its associated label $y \in \mathcal{Y}$, the adversary again tries each possible value $s_i \in \mathcal{S}$ of the sensitive attribute, queries the target model with $(\mathbf{x}, s_i)$, and evaluates how confident the model is in predicting the correct label. The sensitive attribute value that yields the highest confidence for the true label y is inferred as the most likely candidate. This procedure is formalized in Attack 8.

Attack 8 (Confidence Score-Based Model Inversion Attack (CSMIA) [66]) *Let $\mathcal{M}_D$ be a classification model trained on dataset D that outputs a predicted class label y and a confidence vector $\mathbf{c}_i = \mathcal{M}_D(z_i = (\mathbf{x}, s_i))$. The adversary is given the public features $\mathbf{x}$ and the true class label y of a target record $z = (\mathbf{x}, s)$. The attack proceeds as follows:*

- *For each sensitive attribute value $s_i \in \mathcal{S}$, construct the input $(\mathbf{x}, s_i)$ and query the model to obtain the predicted label $\hat{y}_i$ and the corresponding confidence score $\mathbf{c}_i$.*

- *Predict:*

$$\widetilde{s} = \begin{cases} s_i \in \mathcal{S} \mid \hat{y}_i = y, & \text{if exactly one such candidate,} \\ \arg\min_{s_i \in \mathcal{S}} \mathbf{c}_i, & \text{if no prediction matches } y, \\ \arg\max_{s_i \in \mathcal{S},\, \hat{y}_i = y} \mathbf{c}_i, & \text{otherwise.} \end{cases}$$

Finally, Jayaraman et al. [47] introduce the Weighted Confidence-based Attribute Inference (WCAI) attack, a refined black-box method that empirically outperforms both Yeom and CSMIA attacks. Rather than relying solely on membership decisions or raw confidence scores, WCAI combines the target model's confidence with an external imputation model's conditional probabilities to improve inference accuracy.

Again, the adversary constructs inputs $(\mathbf{x}, s_i)$, queries the model $\mathcal{M}_D$ and obtains the confidence scores $\mathbf{c}_i$. The adversary weighs this score by the conditional probability $\Pr[s_i \mid \mathbf{x}]$ estimated from an external imputation model $\mathcal{I}$ trained on auxiliary data. The final inferred sensitive value $\widetilde{s}$ is the candidate maximizing the weighted product of model confidence and imputation probability, as formalized in Attack 9.

Attack 9 (Weighted Confidence-based Attribute Inference (WCAI) [47]) *Given black-box access to a classification model $\mathcal{M}_D$ and the non-sensitive features $\mathbf{x}$ of a target record $z = (\mathbf{x}, s)$, the adversary:*

- *For each candidate sensitive value $s_i \in \mathcal{S}$, constructs the input (x, s_i) and queries the model to obtain the confidence score $\mathbf{c}_i$.*

- *Uses an external imputation model $\mathcal{I} : \mathcal{X} \to \mathcal{S}$ to obtain the conditional probability* $\Pr[s_i \mid \mathbf{x}]$.
- *Predicts the sensitive attribute as*

$$\widetilde{s} = \arg\max_{s_i \in \mathcal{S}} \mathbf{c}_i \cdot \Pr[s_i \mid \mathbf{x}].$$

7.3. Database Descriptions

We use datasets chosen by the authors whose attacks we replicate to maintain consistency and comparability. For the white-box attacks, we use MNIST as in [42], and add the Fashion MNIST dataset to have a second dataset for evaluation. For the black-box and imputation attacks, we use the Census and Texas-100X datasets following [47]. These datasets are publicly available and widely used benchmarks, e.g. [84, 77, 42, 47]. As our focus is on evaluating privacy attacks and metrics rather than advancing model performance, we follow the original dataset choices. We briefly introduce each dataset below.

MNIST: The MNIST dataset [57] is a benchmark collection of 70,000 grayscale images of handwritten digits (0–9). It is split into 60,000 training and 10,000 test examples. To allow for a principled comparison with the results of Hayes et al., we follow their data preprocessing [42].

To construct the database D_-, we load the first 999 training samples. We then load the next $m \in \{8, 128\}$ training samples to form the set $z_1, \ldots, z_m$ of reconstruction candidates. This candidate set serves both as the universe N from which the target z is drawn and as the set of possible reconstructions m, reflecting the adversary's background knowledge.

The task of the target model $\mathcal{M}_D$ is to predict which digit was written, a 10-class classification problem. The attack task is to infer which record from the candidate set $z_1, \ldots, z_m$ was used in training.

Fashion MNIST: The Fashion MNIST dataset [92], referred to as Fashion for short, consists of 70,000 grayscale images of clothing items such as shirts, trousers, and shoes. It is divided into 60,000 training and 10,000 test examples. Compared to MNIST, Fashion MNIST contains more diverse and complex visual patterns. As with MNIST, we load the first 999 training samples to construct the database D_- and the next $m = 8$ samples to form the candidate set $z_1, \ldots, z_m$. This set defines both the universe size N and the number of possible reconstructions m.

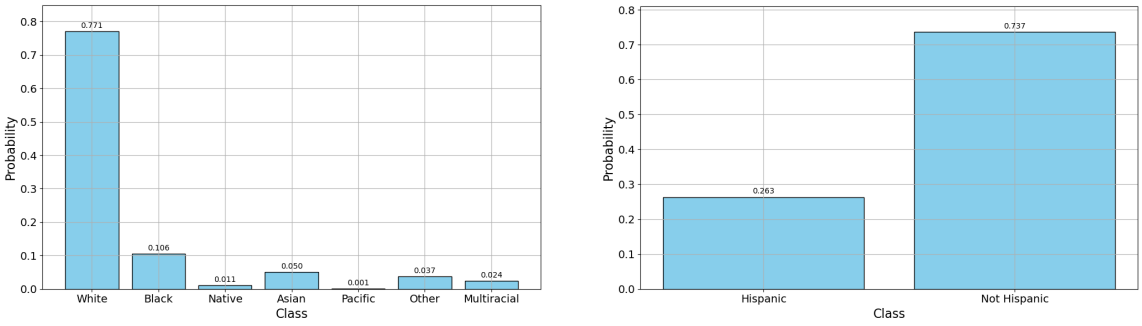
The task of the target model $\mathcal{M}_D$ is to predict which item of clothing is displayed on a given image, a 10-class classification problem. The attack task is the same as on the MNIST dataset.

Census: The Census dataset is based on the 2019 US census and follows the structure of the popular Adult dataset [8]. It contains 1,676,013 records with 14 attributes denoting personal and demographic information about each participant, such as age, gender, race or education.

50,000 records are randomly selected to form the training dataset. From these, we fix 49,000 records as a stable dataset D_- for all experiments, while the remaining 1,000 form a

universe of size N from which target records z are sampled. In the original setting [46], all records were considered part of the training database. For evaluating ReRo and U-ReRo, it is necessary to distinguish a fixed dataset D_- from the universe of target records. We set aside 1,000 records for this purpose, ensuring that the training set of the target model remains sufficiently large.

The target model $\mathcal{M}_D$ is trained to predict whether an individual earns more than 90,000$ per year based on their demographic features. The goal of the attribute inference attacks is to recover the sensitive attribute. Following the setup of Jayaraman et al., we treat race as the sensitive attribute, which constitutes an seven-class classification problem. The possible race categories are: White, Black, Native American, Asian, Pacific Islander, Other Race, and Two or More Races. The distribution of these categories is shown in Figure 7.1.



(a) Race (Census).

(b) Ethnicity (Texas-100X).

Figure 7.1.: Distribution over the sensitive attributes for Census and Texas-100X datasets.

Texas-100X: The Texas-100X dataset is based on the Texas-100 hospital dataset from [84]. A record corresponds to a hospital patient and contains demographic information such as age, gender or ethnicity, and medical information, e.g. the duration of the hospital stay or the diagnosis. The target model $\mathcal{M}_D$ is trained to predict one of 100 surgical procedures based on the patient's record, a 100-class classification problem. Jayaraman et al. curate a new version, Texas-100X, of this dataset that contains 925,128 patient records from 441 hospitals and retains more demographic and medical attributes than the original [47]. We hence work on the newer Texas-100X dataset. The train and test splits are done equivalent to the Census dataset.

For the Texas-100X dataset, ethnicity is the sensitive attribute. More precisely, the task of the adversary is to predict whether an individual is Hispanic or not given their demographic information – a binary classification task. The distribution over the sensitive attribute is shown in Figure 7.1.

7.4. Experimental Design

We use Python and the ML library Tensorflow [2] to evaluate the attacks. However, we do not implement them from scratch but utilize pre-existing code. For the white-box attack,

Database	N	m
MNIST	{8,128}	{8,128}
Fashion MNIST	8	8
Census	1,000	7
Texas-100X	1,000	2

Table 7.1.: Database overview: size of universe N, number of reconstruction candidates m.

Jamie Hayes [42] kindly provided us with a minimal implementation of the attack in a Google Colab, based on which we created our own scripts to extend their experiments to incorporate U-ReRo.

For the imputation and black-box attacks, we adapt the code implementation from the Bargav Jayaraman's public repository: `https://github.com/bargavj/EvaluatingDPML/tree/master/improved_ai` [47].

7.4.1. Computation of ReRo and U-ReRo

The computation of ReRo follows the general procedure outlined in Section 5.5, with some specifics for the ML setting. Each attack requires a target model $\mathcal{M}_D$ trained with a DP mechanism on a dataset D composed of a fixed subset D_- and a target record z. In each iteration, we draw a target record z, add it to D_- to obtain $D = D_- \cup \{z\}$ and then train J models based on this dataset. We evaluate the given attack against each of the J models and average over the results to obtain the ReRo value for this z. The final ReRo value is obtained as the average attack success across all samples z.

To compute the correction term, for each trained model we sample I records z' and evaluate the attack based on $\mathcal{M}_D = \mathcal{M}_{D_- \cup \{z\}}$ against each record z'. We average the attack success over all iterations to estimate the correction term. Finally, U-ReRo is obtained by subtracting this correction term from the previously computed ReRo value.

For the ML attacks, the choice of I and J is driven by both the domain size m and computational feasibility.

For the white-box attacks, the domain size m is defined by the candidate set $\{z_1, ..., z_m\}$ from which the adversary has to choose the correct reconstruction. Across our experiments, we vary this set size. For the standard MNIST experiments, we set $m = 8$ in accordance with [42]. Since the domain size m is relatively small, we can exhaustively iterate over all reconstruction candidates by setting $I = m$. The parameter J specifies how many models we train per reconstruction candidate. For MNIST, it is computationally feasible to train $J = 40$ models. However, when using a larger prior set with $m = 128$, we reduce the number of models to $J = 3$ in order to keep the experiments computationally tractable and results readily available. In the case of the Fashion-MNIST dataset, the model architecture is more complex and requires longer training times, which necessitated reducing the number of models per sample to $J = 10$ to maintain feasibility.

For the black-box attacks, training the models is computationally slow despite GPU acceleration. To balance a sufficient sample size with computational feasibility, we set

$I = 100$ targets and train $J = 5$ models for each. A complete overview of all parameter choices is provided in Table 7.2.

Attack	I	J
White-box	MNIST: {8,128} Fashion: 8	MNIST: {40,3} Fashion: 10
Black-box	100	5

Table 7.2.: Sample sizes used for different attacks. I: number of samples, J: number of models per sample

7.4.2. Upper Bounds

As for the LDP attacks, we only plot the most specific bounds for each setting, as listed in Table 7.3.

Attack	π	η	*aux*	Bound
White-box	$U[m]$ π	0	–	Corollary 2, Corollary 1 Theorem 7, Theorem 8
Imputation	π	0	x	Theorem 9
Black-box	π	0	x	Theorem 9

Table 7.3.: Theoretical bounds for each attack setting.

7.4.3. Target Model Architectures

Attacks in private learning are carried out against a target model. We now describe the different target model architectures we use across our experiments.

White-box attack: For the white-box attack, we use the same model architecture as proposed by Hayes et al. for the MNIST experiments. This is a standard multilayer perceptron that takes a 28×28 grayscale image as input, flattened into a 784-dimensional vector. The input passes through a fully connected layer with 10 hidden units, followed by a ReLU activation. The output layer consists of 10 units, one for each digit class (0–9). For Fashion MNIST, we adapt the model to better capture the spatial structure of the images by using a lightweight convolutional neural network (CNN) [74]. The CNN consists of two convolutional layers: the first has 32 filters with a 3×3 kernel, followed by an ELU activation and a 2×2 max pooling layer. The second convolutional layer has 64 filters, again followed by ELU activation and max pooling. The resulting feature maps are flattened and passed through a fully connected layer with 128 units and ELU activation, before the final output layer with 10 units. The target model is trained with DP-SGD with privacy

parameters $\varepsilon \in \{1, 10, 20, 50\}$ and $\delta = 10^{-5}$. We report the test accuracies of the target model in Table 7.4. All training parameters are taken directly from [42] and can be found in Table 7.5.

ε	**MNIST**	**Fashion**
1	0.46	0.46
10	0.75	0.75
20	0.75	0.77
50	0.75	0.80

Table 7.4.: Average test accuracy for different ε values under DP-SGD on MNIST and Fashion.

Parameter	**Meaning**	**Values**
ε	Privacy budget	$\{1, 5, 10, 20, 50\}$
C	Norm clip	0.1
T	Number of training steps	100
r	Learning rate	1.0 for MNIST, 0.5 for Fashion
m	Size of the prior	8
$\|B\|$	Batch size	1000
q	Batch-sampling rate	1.0
δ	DP parameter	$1e^{-5}$
I	Number of repetitions	100

Table 7.5.: Parameter settings for training the white-box target model.

Black-box attack: For the black-box attacks, we follow the model architecture used by Jayaraman et al. The target model is a simple neural network trained on 49,000 randomly sampled records, with an additional 25,000 records reserved for testing. It consists of two hidden layers with 256 ReLU-activated neurons each, followed by a softmax output layer with one neuron per class (i.e., 2 output neurons for the Census dataset and 100 for the Texas-100X dataset). Models are trained using DP-SGD with privacy parameters $\varepsilon \in 1, 50$ and $\delta = 10^{-5}$. We note that our target models achieve the same accuracies as reported by Jayaraman et al., listed in Table 7.6.

ε	**Census**	**Texas-100X**
1	0.86	0.28
50	0.86	0.38

Table 7.6.: Target model mean test accuracy.

The attacks assume that the adversary has knowledge of the data distribution π. Jayaraman et al. model this by providing the adversary with an auxiliary dataset D_{aux} consisting of records sampled from π, and they vary the size of this dataset to reflect different levels of adversarial knowledge. ReRo and U-ReRo build on the notion of the *informed adversary* where the adversary has full knowledge of D_{-} to carry out their attack [10, 37]. Jayaraman et al. follow a similar approach: For their most powerful setting, the adversary knows all training records except the attack targets [47]. Hence, we set $D_{aux} = D_{-}$ for our attack.

All other target model training parameters are adopted from [47] and summarized in Table 7.7.

Parameter	Meaning	Values
ε	Privacy budget	1
C	Norm clip	4
T	Number of training steps	30
r	Learning rate	0.001
$\|B\|$	Batch size	500
q	Batch-sampling rate	0.01
δ	DP parameter	$1e^{-5}$

Table 7.7.: Parameter settings for training the black-box target model.

7.4.4. Data Distribution for White-box Attack

In the original setting for the white-box attack, every image in the candidate set is considered equally likely. As we have separate bounds for uniform and non-uniform data distributions, we extend this attack to the non-uniform prior setting. This requires defining a data distribution over the images in the candidate set.

To test the tightness of our bounds, we constructed a worst-case distribution. The white-box attack assigns a score to each image in the candidate set based on a given input image (also drawn from the set). This score reflects how likely any other image is to have been the input – essentially a measure of similarity. Using these scores, we first identify the image with the highest overall score which is the most likely to have been the input. Then, for this image, we find the image with the lowest score relative to it, yielding the most distinguishable pair for the attack. We skew the data distribution toward this pair by assigning them a combined probability of over 0.5, split evenly (specifically, 0.35 each for a candidate set of size $m = 8$).

8. Results on Private Learning

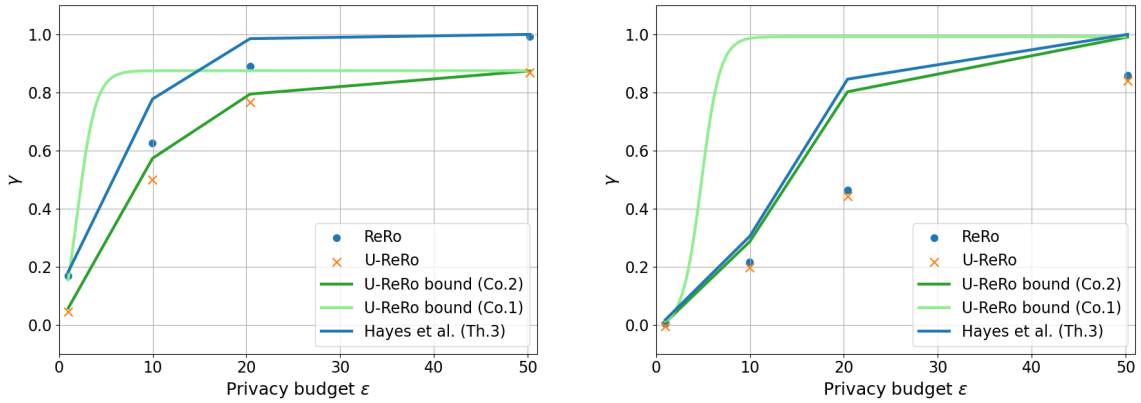
In this chapter, we present the results for the white-box attack and the black-box and imputation attacks.

8.1. Results for the White-box Attack

The results for the white-box attack under a uniform data distribution are shown in Figure 8.1 for the MNIST and in Figure 8.2 for the Fashion dataset. For MNIST, ReRo closely tracks the upper bound reported in [42]. While the attack is more effective on the original MNIST dataset, similar trends are observed for the Fashion dataset.

The upper bound for ReRo is taken directly from [42]. We can confirm that we closely replicate their published results concerning ReRo.

The U-ReRo values are consistently lower than ReRo, due to the candidate set size of $m \in 8, 128$. As a baseline, an adversary has a $\frac{1}{m} \in 0.125, 0.01$ chance of correctly guessing the reconstruction at random. This baseline is explicitly accounted for in the U-ReRo correction term, which reduces the final U-ReRo value. To verify that the gap between ReRo and U-ReRo arises from this random guessing baseline, we repeat the experiment with a larger prior set of size $m = 128$. The results in Figure 8.1 show that ReRo and U-ReRo align closely, with a correction term of $\frac{1}{128} \approx 0.007$.



(a) Candidate set size $m = 8$.

(b) Candidate set size $m = 128$.

Figure 8.1.: U-ReRo bound and attack performance on DP-SGD trained model (MNIST dataset) for different candidate set sizes and uniform prior.

As the correction term only accounts for random guessing, the attack is genuinely effective – privacy leakage only occurs when an image is actually included in the training

dataset. For images drawn from the underlying data distribution, the attack performs no better than random guessing. To understand why, we consider how the attack works: the adversary observes all gradient updates from training on $D_- \cup \{z\}$. They subtract the known gradients from D_- to isolate those attributable to z. However, in the case of U-ReRo, these isolated gradients are then used to compute a loss for a *different* target z_1. As a result, the gradients provide no meaningful information about z_1, and the adversary has no advantage over random guessing.

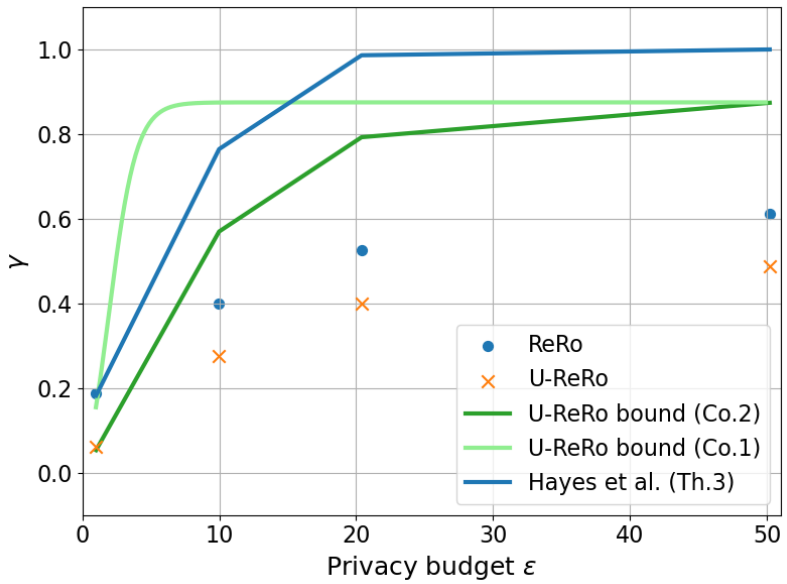
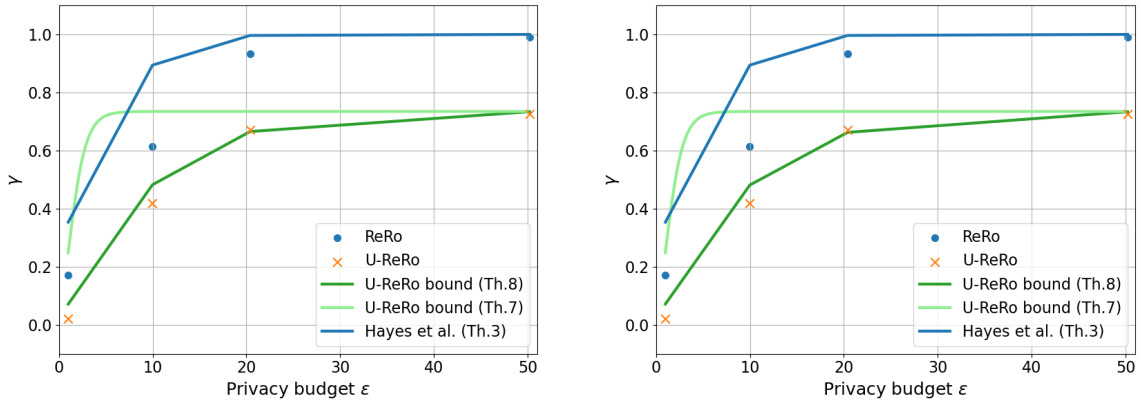


Figure 8.2.: U-ReRo bound and attack performance on DP-SGD trained model (Fashion dataset) for candidate set size $m = 8$ under uniform prior.

Next, we evaluate the attack under a non-uniform data distribution–the prior distribution defined in Section 7.4.4. The results for both datasets are shown in Figure 8.3. Compared to the uniform prior setting, the gap between ReRo and U-ReRo is noticeably larger. At higher privacy budgets, the U-ReRo correction term converges to 0.35, which matches the maximum candidate probability under this data distribution. This occurs because, in this setting, the likelihood of the target image coinciding with a training sample is substantially higher than for the uniform prior setting, which in turn increases the baseline success probability.

Regarding the tightness of the bounds, we observe that our bounds from Corollary 2 and Theorem 8 are perfectly tight for both the uniform and non-uniform prior setting on the MNIST dataset for $m = 8$. This corresponds to the setting where Hayes et al. also demonstrated the near-tightness of their own ReRo-bound. We note that the attack performs less well on the Fashion dataset and for the larger candidate set size $m = 128$. However, since we know that the attack is near-optimal [42] and that we see its performance match the bound exactly for $m = 8$ on the MNIST dataset, we can conclude that our bounds neither over- nor underestimate the privacy leakage for this setting but accurately reflect the risk.



(a) MNIST ($m = 8$). (b) Fashion ($m = 8$).

Figure 8.3.: U-ReRo bound and attack performance on DP-SGD trained model under non-uniform prior (MNIST and Fashion datasets).

While the general bounds from Corollary 1 and Theorem 7 are typically looser–since they apply to any DP mechanism rather than being tailored to DP-SGD–they still outperform existing results in certain settings. For example, with a candidate set size of $m = 8$, our general bound is tighter than the ReRo bound from [42] for all $\varepsilon \geq 15$ under the uniform prior and for all $\varepsilon \geq 9$ under the non-uniform prior. This demonstrates that our bound can provide a more accurate risk assessment even when the underlying DP mechanism is unspecified.

8.2. Results for the Imputation and Black-box Attacks

The results for the imputation and black-box attacks are reported in Table 8.1 for the Census dataset and in Table 8.2 for the Texas-100X dataset, respectively.

Examining the utility of the target model trained with DP-SGD, we see that for the Census dataset the test accuracy remains stable around 0.86 for all $\varepsilon \in 1, 50$, cf. Table 7.6. This aligns with the results reported by Jayaraman et al., indicating that DP-SGD, in this setting, does not noticeably degrade model performance. Notably, our test accuracy for $\varepsilon = 1$ matches their reported value exactly, confirming the correctness of our setup. We note that the original authors only published results for $\varepsilon = 1$ [47].

Given this stability in utility, we do not expect major variations in attack performance under different privacy budgets–and the results in Table 8.1 support this expectation. Across privacy budgets, the attack performance remains largely constant. The imputation attack is independent of the target model which is reflected by the constant U-ReRo values across privacy budget. The black-box attacks–Yeom, CSMIA, and WCAI–also show little to no variation, consistent with the unchanged model accuracy. The largest fluctuation appears in WCAI, with Aux-Aware ReRo dropping from 0.86 to 0.79. Because this attack was computationally costly and we had to reduce the sample size to $J = 100$, we attribute the drop mainly to sampling error. This interpretation is supported by the fact that Aux-

Aware U-ReRo also decreases (from 0.08 to 0), suggesting that little leakage was present for $\varepsilon = 1$ in the first place.

More importantly, we note that all attacks achieve consistently high ReRo scores–both the imputation baseline and the black-box attacks. In contrast, U-ReRo values remain very low. This strongly suggests that the attacks are not exploiting training set participation, but instead leverage patterns inherent in the data distribution. These findings are consistent with Jayaraman et al., who likewise observe that simple imputation performs as well as, or better than, black-box attacks when the adversary has substantial background knowledge D_{aux} [47].

ε	**Imputation**	**Yeom**	**WCAI**	**CSMIA**
1	ReRo: 0.82 U: 0	ReRo: 0.8 U: 0.05	ReRo: 0.86 U: 0.08	ReRo: 0.81 U: 0.04
50	ReRo: 0.82 U: 0	ReRo: 0.8 U: 0.03	ReRo: 0.79 U: 0	ReRo: 0.83 U: 0.05

Table 8.1.: ReRo and U-ReRo for black-box ML attacks on the Census dataset. Test accuracy of the target model is 0.857 for $\varepsilon = 1$ and 0.863 for $\varepsilon = 50$.

Turning to the Texas-100X dataset, we observe a different pattern compared to Census. With DP-SGD, the target model suffers a substantial loss in accuracy: as shown in Table 7.6, test accuracy drops to 0.28 for $\varepsilon = 1$ and recovers only slightly to 0.38 for $\varepsilon = 50$. This indicates that, unlike in the Census setting, DP here clearly reduces model utility. Even with a high privacy budget, however, the model performs poorly with an overall test accuracy of only 0.38. Our model accuracy for $\varepsilon = 1$ is consistent with the results from [47].

Despite this degradation in accuracy, the imputation baseline remains constant, as expected, highlighting that this attack is model-independent. Moreover, the imputation attack in this setting outperforms all black-box attacks by a wide margin. ReRo yields 0.71 for both privacy budget whereas U-ReRo is close to zero, as expected.

For the black-box attacks, we observe a substantial drop in ReRo values when moving from $\varepsilon = 1$ to $\varepsilon = 50$. In particular, U-ReRo even turns negative. We attribute these irregularities to several factors. First, the target model's test accuracy is low (cf. Table 7.6), which makes the attack results more unstable. Second, compared to the Census dataset, attacks appear more difficult on Texas-100X as overall ReRo values are lower. Combined with our limited sample size, this likely amplifies sampling error. Negative U-ReRo values in particular suggest that too few samples were available to obtain a reliable estimate. Taken together, these factors explain the noisy outcomes and indicate that attack performance here is shaped more by data and model limitations than by the privacy budget.

Nevertheless, we observe the same overall trend than on the Census dataset: ReRo values are high and U-ReRo values very low, not only for the imputation attack but also for each of the tested black-box attacks. Attack success is then not primarily attributable to participation in the target model's training dataset. This finding further strengthens the conclusion that black-box attack success does not imply privacy leakage in this setting, in line with the observations from Jayaraman et al.

ε	Imputation	Yeom	WCAI	CSMIA
1	ReRo: 0.71 U: -0.01	ReRo: 0.54 U: 0.06	ReRo: 0.58 U: 0.08	ReRo: 0.3 U: 0.01
50	ReRo: 0.71 U: 0	ReRo: 0.4 U: -0.07	ReRo: 0.41 U: -0.06	ReRo: 0.26 U: -0.01

Table 8.2.: ReRo and U-ReRo for black-box ML attacks on the Texas-100X dataset. Test accuracy of the target model is 0.273 for $\varepsilon = 1$ and 0.376 for $\varepsilon = 50$.

Overall, our evaluation confirms the findings of Jayaraman et al.: the black-box attacks under analysis do not pose a significant privacy risk beyond what can already be achieved via simple imputation, in this setting. While Jayaraman et al. demonstrated this through comparisons with an imputation model, our results show that this conclusion is directly supported by the U-ReRo metric. Unlike its predecessor ReRo–which here overestimates attack effectiveness by conflating data correlation with leakage-U-ReRo correctly reveals that none of the attacks, including imputation, exploit information that resulted from records being present in the target model's training dataset. In both datasets, U-ReRo values remain consistently near zero, indicating that successful inferences arise from patterns in the underlying data distribution, not from participation in the training set.

8.3. Intermediate Conclusions

Our evaluation of the state-of-the-art white-box attack by Hayes et al. confirms that their method genuinely exploits private information resulting from the target's participation in the training dataset, as revealed through intermittent gradients. This is evidenced by the close alignment between ReRo and its theoretical upper bound, indicating that the attack is near-optimal in the evaluated setting. We further support this conclusion with our U-ReRo metric, which in this case only discounts random guessing and thus demonstrates that the attack extracts participation-specific information beyond what could be inferred from the background knowledge alone.

Our bound on U-ReRo under f-DP are perfectly tight for this setting, both for uniform and non-uniform data distributions. We significantly improve over the state-of-the-art white-box bound from Hayes et al. (Theorem 3). Even our black-box bounds under (ε, δ)-DP that are not specific to DP-SGD still improve upon their bound for mid-range and high privacy budgets.

Our evaluation of black-box attacks and data imputation reveals a different pattern. Data imputation models, which rely purely on correlations within the data, can achieve high ReRo values when predicting sensitive attributes, leading to an overestimation of privacy risks. However, since these models do not extract information related to participation, Aux-Aware U-ReRo remains zero–indicating no actual privacy leakage.

Furthermore, our experiments confirm the findings of Jayaraman et al., as we did not observe any black-box attacks which yielded Aux-Aware U-ReRo values significantly above zero. This supports the conclusion that such attacks, while possibly effective at identifying

population-level patterns, do not succeed in exploiting private information about specific individuals.

Part III.

Analyzing U-ReRo for DP Auditing

9. Introduction

DP offers strong, worst-case theoretical guarantees against individual privacy leakage [29]. However, these guarantees do not always reflect how mechanisms behave in complex, real-world deployments. DP auditing has therefore emerged as a complementary approach, empirically evaluating whether deployed mechanisms fulfill their privacy guarantees in practice [5, 45]. Unlike theoretical ε values, which are derived from worst-case analyses, empirical ε values are computed from observed adversarial success rates and give a perspective on real-world risk. Together, formal guarantees and empirical audits provide a comprehensive understanding of privacy risks.

Current auditing approaches predominantly focus on membership inference attacks. In many real-world scenarios, membership is either publicly known or not inherently sensitive [10], and adversaries may instead turn to attribute inference or data reconstruction. This creates a gap in our auditing capabilities: we lack tools to assess privacy risks in settings beyond membership inference. As a result, current audits may give an incomplete understanding of privacy risk.

To address this gap, we present a novel empirical auditing framework based on the U-ReRo metric. By quantifying how accurately adversaries can infer sensitive information about their target–and doing so with a tight connection to the DP privacy budget ε–U-ReRo enables audits based on any attack strategy.

We evaluate our approach in the setting of LDP. While some work in the auditing literature explores strategies that rely on very few runs of the mechanism [85, 61], such methods are not directly applicable to U-ReRo, which depends on repeated evaluation of the DP mechanism. We instead follow the line of work that audits by running the DP mechanism many times [45, 6]. LDP mechanisms are computationally lightweight, making them well-suited for this setting.

The LDP context also allows for a direct comparison with the state-of-the-art tool LDP Auditor [6], providing a concrete benchmark to assess the effectiveness of U-ReRo as an auditing metric. We follow the assumptions of LDP Auditor and perform audits in a uniform prior setting. To enable a fair comparison with the state-of-the-art, we employ our tightest bound under uniform prior (Corollary 1) on classic U-ReRo for the auditing experiments.

9.1. U-ReRo-based DP Auditing

The core idea of U-ReRo-based auditing is simple: given a measured U-ReRo value, we invert our theoretical bound to estimate the corresponding privacy budget. This yields an empirical ε, reflecting the level of privacy loss observed in practice.

In order to obtain an estimate for the empirically observed privacy budget, denoted as $\widetilde{\varepsilon}$, from a given U-ReRo score $\widetilde{\gamma}$, we make use of the theoretical relationship between γ and the true privacy budget ε. In earlier chapters, we derived an upper bound on the U-ReRo value γ as a function of the privacy parameter ε, capturing the worst-case adversarial success under DP. We can write this relationship as:

$$\gamma \leq f(\varepsilon),$$

where $f(\varepsilon)$ is a function that characterizes the maximum possible U-ReRo score under a given privacy budget ε. To estimate the empirical privacy loss $\widetilde{\varepsilon}$ from an observed U-ReRo score $\widetilde{\gamma}$, we invert the function f to obtain:

$$\widetilde{\varepsilon} = f^{-1}(\widetilde{\gamma}).$$

This inversion allows us to interpret the observed adversarial performance (quantified by $\widetilde{\gamma}$) in terms of an equivalent privacy budget, yielding an estimate of the actual privacy leakage in practice.

It should be noted that this approach implicitly assumes that the observed U-ReRo score lies within the range where f is invertible and meaningful. We shall see based on the exact formulas of the various theoretical bounds that this may not always be the case.

The quality of this estimate depends on the quality of the attack that yields $\widetilde{\gamma}$. If the attack is strong and well-optimized–meaning it closely approaches the theoretical maximum defined by $f(\varepsilon)$–then the estimated privacy budget $\widetilde{\varepsilon}$ should be close to the true value. In this case, the inversion is reliable, and we obtain a tight, meaningful estimate of the actual privacy leakage observed in practice.

Conversely, if the attack is weak, the observed $\widetilde{\gamma}$ may fall significantly below the theoretical maximum. In such cases, the inverted estimate $\widetilde{\varepsilon}$ will underestimate the true privacy leakage, since the adversary has not fully exploited the available information. Empirical privacy estimation lower-bounds the true ε, with tighter bounds for stronger attacks.

Furthermore, if the estimated privacy budget $\widetilde{\varepsilon}$ exceeds the theoretical privacy budget ε, this may signal a violation of the DP assumptions, such as an incorrect implementation of the privacy mechanism. Such discrepancies can serve as a powerful diagnostic tool for identifying bugs and other DP violations in practice [14].

We present the formulas to derive the empirical privacy budget for $(0, \gamma)$-U-ReRo under uniform prior in Method 1.

Method 1 (Empirical ε for $(0, \gamma)$-U-ReRo under uniform prior (based on Corollary 1)) *Let m follow Corollary 1 and let $\widetilde{\gamma}$ be the result of the U-ReRo metric. From Corollary 1, we know that U-ReRo is upper-bounded by*

$$\frac{\mathrm{e}^{\varepsilon} - 1}{\mathrm{e}^{\varepsilon} + m - 1} \frac{m-1}{m}.$$

We then derive the empirical epsilon value $\widetilde{\varepsilon}$ as follows:

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

Justification: The empirical $\widetilde{\varepsilon}$ is obtained by inverting the uniform-prior bound

$$\widetilde{\gamma} \leq \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + m - 1} \frac{m-1}{m}.$$

Multiplying both sides by $\frac{m}{m-1}$ and rearranging terms gives

$$e^{\varepsilon} \left(1 - \widetilde{\gamma} \frac{m}{m-1}\right) \geq \widetilde{\gamma} m + 1,$$

so that

$$\widetilde{\varepsilon} = \ln \left(\frac{\widetilde{\gamma} m + 1}{1 - \widetilde{\gamma} \frac{m}{m-1}} \right).$$

9.2. Benchmark: LDP AUDITOR

We briefly describe LDP AUDITOR [6], the tool we will use as a benchmark. It estimates the empirical privacy loss $\widetilde{\varepsilon}$ of a given LDP mechanism by evaluating how accurately an adversary can distinguish between two input values v_1 and v_2 after applying the mechanism. Over multiple trials, it measures the number of times the adversary correctly identifies v_1 from the output of $\mathcal{M}(v_1)$ (true positives) and incorrectly identifies v_1 from $\mathcal{M}(v_2)$ (false positives). Confidence intervals for these rates are computed using the Clopper–Pearson method, and the resulting bounds are used to estimate $\widetilde{\varepsilon}$. The algorithmic description can be found in Algorithm 1.

Algorithm 1 LDP Auditor [6]

Require: Theoretical ϵ-LDP protocol $\mathcal{M}$, attack A, values $v_1, v_2 \in \mathcal{V}$, trial count T, confidence level α
Ensure: Estimated empirical privacy loss $\widetilde{\varepsilon}$

$TP \leftarrow 0,\ FP \leftarrow 0$ ▷ True Positives and False Positives
for $i = 1$ to T **do**
  if $A(\mathcal{M}(v_1)) = v_1$ **then**
    $TP \leftarrow TP + 1$
  end if
  if $A(\mathcal{M}(v_2)) = v_1$ **then**
    $FP \leftarrow FP + 1$
  end if
end for
$\hat{p}_0 \leftarrow \text{ClopperPearsonLower}(TP, T, \alpha/2)$
$\hat{p}_1 \leftarrow \text{ClopperPearsonUpper}(FP, T, \alpha/2)$
return $\widetilde{\varepsilon} = \ln((\hat{p}_0)/\hat{p}_1)$

An inherent limitation of LDP AUDITOR is that the choice of parameters T and α restricts the maximum estimable value of ε. In particular, repeating the experiments with $T = 1e6$ and $\alpha = 0.01$ yields a Monte Carlo upper bound of $\widetilde{\varepsilon} = 12.25$ [6]. Beyond this point, the estimates of ε stagnate, and LDP AUDITOR is unable to provide higher values.

9.3. Mechanism Selection

To enable a thorough comparison with LDP Auditor, we adopt the same LDP mechanisms: GRR, Subset Selection (SS), Unary Encoding (UE), Histogram Encoding (HE), and Local Hashing (LH) [6]. These mechanisms are widely studied and serve as core building blocks for more complex LDP protocols [30, 90, 11]. Evaluating on this set ensures relevance to practical applications and allows us to assess auditing performance across a diverse range of mechanisms. We present a brief overview of the mechanisms that have not been discussed before.

SS: In SS [96, 90], users report a subset $\Omega \subseteq \mathcal{Z}$ containing their true value z with probability $p = \frac{\omega e^{\varepsilon}}{\omega e^{\varepsilon}+m-\omega}$, where $\omega = |\Omega| = \max\left(1, \left\lfloor \frac{m}{e^{\varepsilon}+1} \right\rfloor\right)$. The subset is completed by sampling uniformly from $\mathcal{Z} \setminus \{z\}$.

UE: UE [30, 90] encodes the user's input $z \in \mathcal{Z}$ as a one-hot m-dimensional binary vector $\mathbf{z}$ and perturbs each bit independently. For each position $i \in [m]$, the obfuscated vector $\boldsymbol{\theta}$ is sampled such that $\Pr[\boldsymbol{\theta}_i = 1] = p$ if $\mathbf{z}_i = 1$, and q otherwise. Two variants exist: (i) *Symmetric UE (SUE)* [30] with $p = \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$ and $q = \frac{1}{e^{\varepsilon/2}+1}$, and (ii) *Optimal UE (OUE)* [90] with $p = \frac{1}{2}$ and $q = \frac{1}{e^{\varepsilon}+1}$.

HE: HE [90] encodes the user's input $z \in \mathcal{Z}$ as a one-hot m-dimensional vector $\mathbf{z}$ with a single entry equal to 1 and all others 0. Then, each component is independently perturbed using the Laplace mechanism: $\boldsymbol{\theta}_i = \mathbf{z}_i + \mathrm{Lap}\left(\frac{2}{\varepsilon}\right)$, where the global sensitivity is $\Delta_1 = 2$ due to the ℓ_1-distance between any two one-hot vectors.

LH: LH [11, 90] uses a family of hash functions $\mathcal{H}$ to map an input $z \in \mathcal{Z}$ into a smaller domain $[g]$, followed by the application of GRR to the hashed output. Each user selects a hash function $H \in \mathcal{H}$ at random and reports $\langle H, \theta \rangle$, where $\theta = \mathcal{M}_{GRR}(H(z))$. Two variants are: (i) Binary LH (BLH) [11] with $g = 2$, and (ii) Optimal LH (OLH) [90] with $g = \lfloor e^{\varepsilon} + 1 \rfloor$.

9.4. Attack Descriptions

We adopt the attack strategies for each LDP protocol directly from [39, 7, 6] to allow for a fair comparison between their auditing results and ours. We briefly introduce each attack, exempting GRR as the attack strategy is the same as in Attack 1.

For the attack on SS, the adversary receives the subset which contains their target's true value and outputs a random value from this subset as a prediction, cf. Attack 10.

Attack 10 (Attack on SS [7, 39]) *Let $\mathcal{M}_{SS}$ be the SS mechanism with domain $\mathcal{Z}$ and subset size ω. Let $z \in \mathcal{Z}$ be the true user value, and $\Omega = \mathcal{M}_{SS}(z) \subseteq \mathcal{Z}$ the reported subset, such that $z \in \Omega$. The adversary:*

1. *Receives the reported subset Ω,*
2. *Outputs a guess $z' \sim U[\Omega]$.*

In the UE protocol, the adversary receives the target's perturbed vector and selects a value uniformly at random from the indices set to one, if any exist; otherwise, they select uniformly from the entire domain. The attack is formally described in Attack 11.

Attack 11 (Attack on UE [7, 39]) *Let $\mathcal{M}_{UE}$ be the UE mechanism with domain $\mathcal{Z}$ of size m. The user's true value $z \in \mathcal{Z}$ is encoded as a one-hot vector $\mathbf{z} \in \{0,1\}^m$ and perturbed to $\boldsymbol{\theta} = \mathcal{M}_{UE}(\mathbf{z})$.*

The adversary:

1. *Receives the perturbed vector $\boldsymbol{\theta} \in \{0,1\}^m$,*
2. *Constructs the candidate subset $\mathbf{1}_{\mathrm{UE}} = \{i \in [m] \mid \boldsymbol{\theta}_i = 1\}$,*
3. *Outputs*

$$z' \sim \begin{cases} U[m] & \text{if } \mathbf{1}_{\mathrm{UE}} = \emptyset, \\ U[\mathbf{1}_{\mathrm{UE}}] & \text{otherwise.} \end{cases}$$

Following Arcolezi et al., we evaluate two proposed attack strategies on the HE mechanism: Summation with Histogram Encoding (SHE) [6, 29] and Thresholding with Histogram Encoding (THE) [90, 6]. In SHE, the adversary uses Bayesian inference: they calculate the likelihoods based on the noisy vector and select the most likely value, cf. Attack 12. For THE, the adversary picks a threshold value τ to which they compare all entries of the noisy vector and then selects uniformly from the indices of those entries above the threshold, as described in Attack 13.

Attack 12 (SHE Attack [29, 6]) *Let $\mathcal{M}_{HE}$ be the HE mechanism with domain $\mathcal{Z}$ of size m. The user's true value $z \in \mathcal{Z}$ is encoded as a one-hot vector $\mathbf{z} \in \{0,1\}^m$ and perturbed by adding Laplace noise to each entry, yielding $\boldsymbol{\theta} = \mathbf{z} + Lap(2/\varepsilon)^m$.*

The adversary:

1. *Receives the noisy vector $\boldsymbol{\theta} \in \mathbb{R}^m$,*
2. *Computes, for each $z \in \mathcal{Z}$, the likelihood*

$$P(\boldsymbol{\theta} \mid \mathbf{z}) = \frac{1}{(2b)^m} \exp\left(-\frac{\|\boldsymbol{\theta} - \mathbf{z}\|_1}{b}\right), \quad \text{with } b = \frac{2}{\varepsilon},$$

3. *Outputs the most likely value :*

$$z' = \arg\max_{z \in \mathcal{Z}} P(\boldsymbol{\theta} \mid \mathbf{z}).$$

Attack 13 (Attack on THE [90, 6]) *Let $\mathcal{M}_{HE}$ be the HE mechanism with domain $\mathcal{Z}$ of size m. The user's true value $z \in \mathcal{Z}$ is encoded as a one-hot vector $\mathbf{z} \in \{0,1\}^m$ and perturbed to $\boldsymbol{\theta} = \mathbf{z} + Lap(2/\varepsilon)^m$.*

The adversary:

1. *Receives the noisy vector $\boldsymbol{\theta} \in \mathbb{R}^m$,*

2. *Constructs a support set* $\mathbf{1}_{\text{THE}} = \{z \in \mathcal{Z} \mid \boldsymbol{\theta}_z > \tau\}$ *for some threshold* $\tau \in (0.5, 1)$,

3. *Outputs a guess* z' *drawn uniformly from* $\mathbf{1}_{\text{THE}}$ *if it is non-empty, or from* $[m]$ *otherwise.*

Finally, for the LH protocol, the adversary receives the user's perturbed hashed output θ as well as the hash function H that was used. They then iterate over all $z \in \mathcal{Z}$ to find those that give the output θ when passed through H. From this set of possible candidates, the adversary then outputs a random guess. The attack is formally described in Attack 14.

Attack 14 (Local Hashing Attack (LH) [39, 7]) *Let* $\mathcal{M}_{LH}$ *be the LH mechanism over domain* $\mathcal{Z}$ *with hash range* $[g]$ *and hash family* $\mathcal{H}$*. The user's true value* $z \in \mathcal{Z}$ *is hashed using a randomly chosen function* $H \in \mathcal{H}$*, and the hash output* $H(z)$ *is perturbed with the GRR mechanism, yielding* $\theta = \mathcal{M}_{GRR}(H(z))$*. The user reports the pair* $\langle H, \theta \rangle$*. The adversary:*

1. *Receives the pair* $\langle H, \theta \rangle$,

2. *Constructs a support set* $\mathbf{1}_{\text{LH}} = \{z \in \mathcal{Z} \mid H(z) = \theta\}$,

3. *Outputs*

$$z' \sim \begin{cases} U[m] & \text{if } \mathbf{1}_{\text{LH}} = \emptyset, \\ U[\mathbf{1}_{\text{LH}}] & \text{otherwise.} \end{cases}$$

9.5. Database Selection

To evaluate U-ReRo for auditing in the LDP setting, we return to the datasets used in our analysis of LDP mechanisms: the Porto and Beijing datasets. A key advantage of these datasets is their significantly larger domain sizes–$m = 3052$ for Porto and $m = 5356$ for Beijing–compared to the relatively small domains ($m \leq 200$) evaluated in LDP Auditor [6]. This choice provides a more challenging auditing scenario: Arcolezi et al. observed that auditing becomes increasingly difficult as domain size grows. By extending the evaluation to large-domain settings, we explore assess U-ReRo as an auditing tool in a challenging setting based on real-world data.

9.6. Experimental Design

The computation of U-ReRo follows the procedure outlined in Section 5.5. Since LDP Auditor relies on 10^6 runs of the mechanism to obtain reliable results [6], we align our evaluation with a comparable sample size. Specifically, we set I to the size of the data domain ($I = 3052$ for Porto and $I = 5356$ for Beijing), and choose $J = 10^6/I$, ensuring the total number of runs matches that of LDP Auditor. Following Arcolezi et al., we furthermore repeat the estimation five times and report both the mean and the standard deviation across these repetitions.

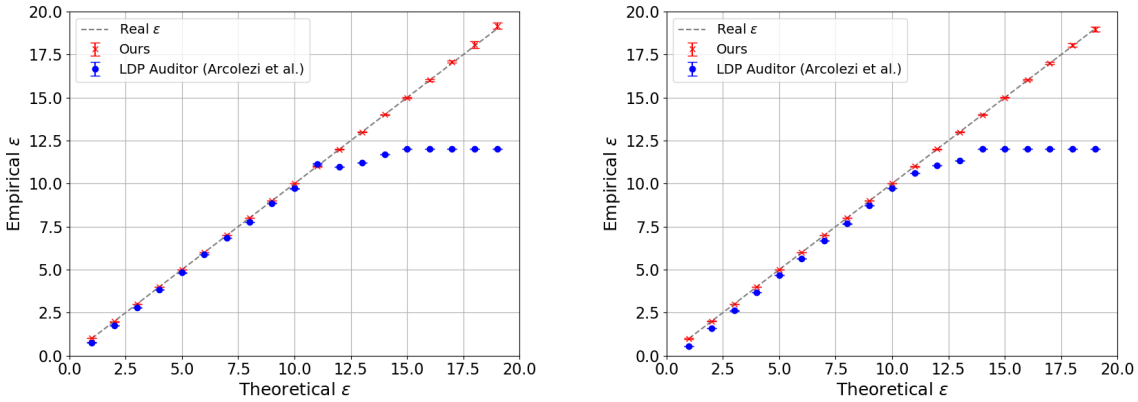
To obtain the results for LDP Auditor, we used the code from Héber Hwang Arcolezi's public GitHub repository (`https://github.com/hharcolezi/ldp-audit`) [6] and ran it for our respective domain sizes of $m \in \{3052, 5356\}$.

10. Results

First, we compare our auditing results to those from LDP Auditor across the various LDP mechanisms. Then, we evaluate U-ReRo-based auditing as a tool to identify bugs in the implementation of LDP mechanisms.

10.1. Comparison to LDP Auditor

The results for the GRR mechanism are shown in Figure 10.1. We recall that our bound from Corollary 1 was perfectly tight for this setting which should yield a strong approximation. Indeed, we observe a perfect alignment between our auditing results and the theoretical privacy budget, indicating that our auditing procedure is tight in this setting. In contrast, the auditing results for LDP Auditor plateau for $\varepsilon \geq 12$ due to its parameter constraint. U-ReRo-based auditing is not limited by these constraints and continues to produce tight estimates. Hence, our auditing procedure yields significantly tighter results for the higher values of the privacy budget.



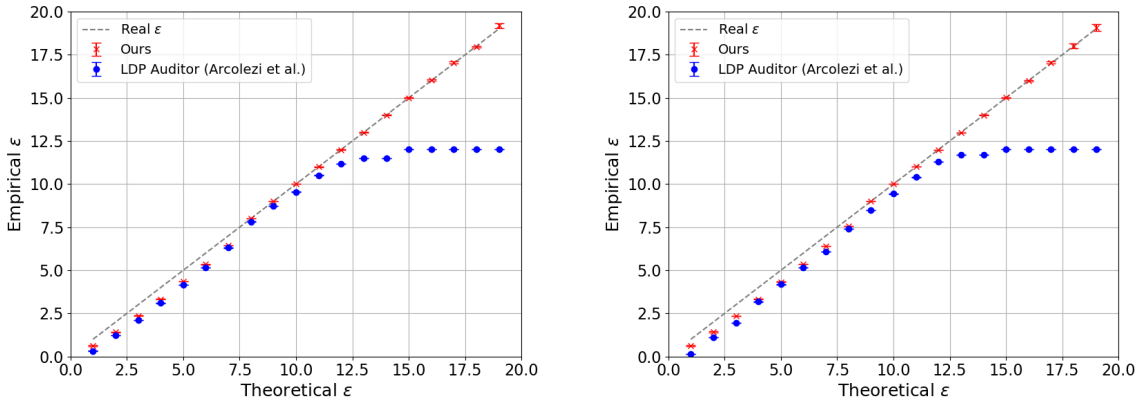
(a) Porto dataset.

(b) Beijing dataset.

Figure 10.1.: Epsilon estimation for Attack 1 on GRR.

We continue with the results for the SS mechanism, shown in Figure 10.2. The results are similar to those observed for GRR, although we consistently find a slight underestimation of the empirical privacy budget for low to mid-range values of $\varepsilon \leq 9$.

The discrepancy in the low-range ε can be attributed to the fundamental difference between the mechanisms: while GRR perturbs individual values directly, SS constructs a subset that includes the true value with a certain probability. This introduces additional variability, resulting in some untightness in the estimation up to the mid-range privacy



(a) Porto dataset.

(b) Beijing dataset.

Figure 10.2.: Epsilon estimation for Attack 10 on SS.

budgets. Nevertheless, the estimated values closely follow the empirical privacy levels and converge around $\varepsilon \approx 10$.

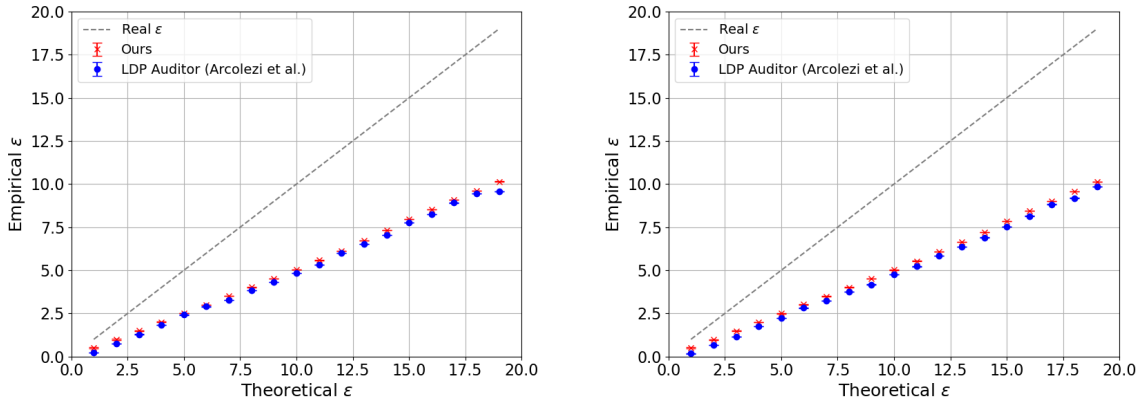
Compared to LDP Auditor, our method achieves comparably for low values of the privacy budget and gives tighter lower bounds on both datasets for the higher privacy budget, as the estimates of LDP Auditor plateau due to its parameter constraints.

Results for the attack on the UE mechanism are plotted in Figure 10.3. As shown in the figures, the attack on SUE performs worse than the one on OUE across lower values of ε. This is expected, as SUE selects its parameters such that $p+q = 1$, leading to symmetric noise and higher variance. In contrast, OUE optimizes its parameters to minimize the estimation error [39], resulting in better utility and more accurate reconstruction, particularly in the low and mid-range privacy budgets.

We furthermore observe that the performance of OUE plateaus as ε increases, consistent with the observations in [6]. This plateau occurs because the adversarial success rate under OUE has an inherent upper bound and converges to $\frac{1}{2}$, as shown in [39]. This limit stems from the fixed parameter $p = \frac{1}{2}$ in OUE, which caps the distinguishability of any input value even as ε grows. Our empirical privacy estimates are comparable with those from LDP Auditor across both datasets and mechanisms.

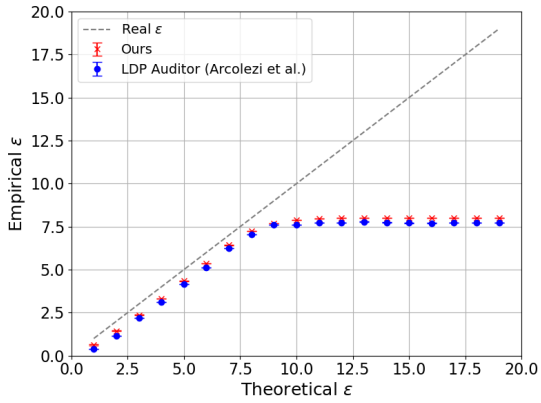
With the SHE and THE attacks, as shown in Figure 10.4, we observe an overall less accurate estimation of the privacy budget. Notably, although the estimated privacy budget increases with the true ε, the gap between the estimated and actual values also grows–especially in the moderate to high privacy budget range for both protocols. We note that we perform comparably to LDP Auditor in both attacks, indicating that the less tight estimate is due to the attack strategy rather than our auditing approach.

Finally, we examine the results for the BLH and OLH mechanisms, shown in Figure 10.5. As noted by Arcolezi et al., the attack on BLH proves largely ineffective in larger data domains. BLH hashes the input set of size m into $\{0, 1\}$, leading to significant information loss–even for the relatively small domain sizes considered in their study [6]. Given that our domain sizes are over ten times larger, the limited attack success and resulting inability to estimate ε accurately is unsurprising.

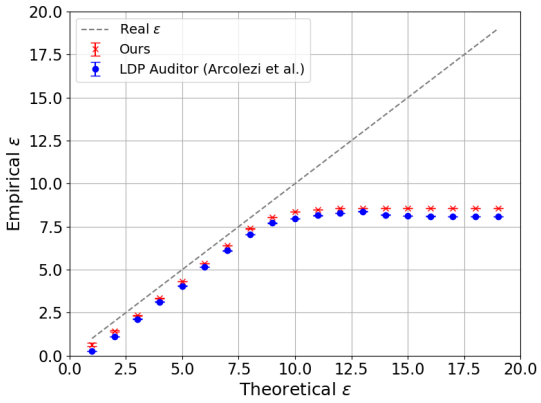


(a) Porto dataset.

(b) Beijing dataset.



(c) Porto dataset.



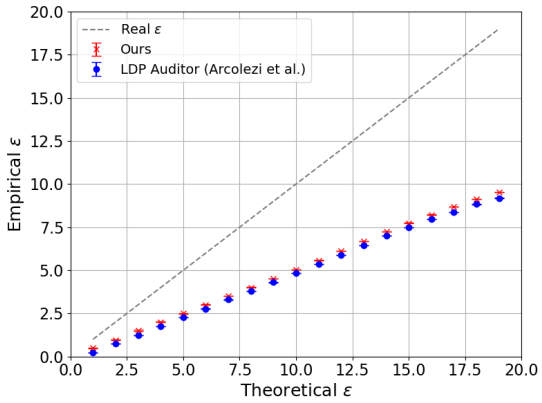
(d) Beijing dataset.

Figure 10.3.: Epsilon estimation for Attack 11 the UE mechanisms: SUE on the top row and OUE on the bottom row.

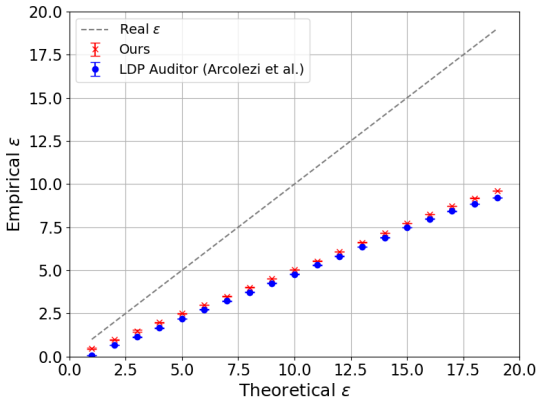
For OLH, we only audit the mechanism until $\varepsilon = 10$ because the output domain size of the hash function grows exponentially with ε which becomes computationally infeasible for large values of the privacy budget. The privacy budget estimates for OLH are considerably tighter than for BLH. For both protocols, we perform comparably to LDP AUDITOR.

Overall, our results demonstrate that U-ReRo performs comparably to and for GRR and SS outperforms LDP AUDITOR, representing an improvement over the current state of the art in LDP auditing. In particular, U-ReRo provides tighter estimates of the empirical privacy budget and remains reliable across a broader range of parameters.

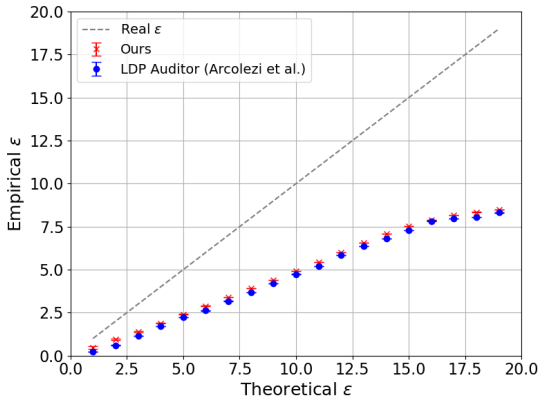
A key advantage of U-ReRo is its flexibility: unlike LDP AUDITOR, it is not tied to indistinguishability attacks but adapts auditing to any attack strategy. For the experiments presented above we employed our bound from Corollary 1 for classic U-ReRo under uniform prior but the same method can be applied for Aux-Aware U-ReRo and its bound – achieving ture generality in the attack strategy. While a broader evaluation across diverse reconstruction attack strategies lies beyond the scope of this thesis, it presents a promising direction for future work.



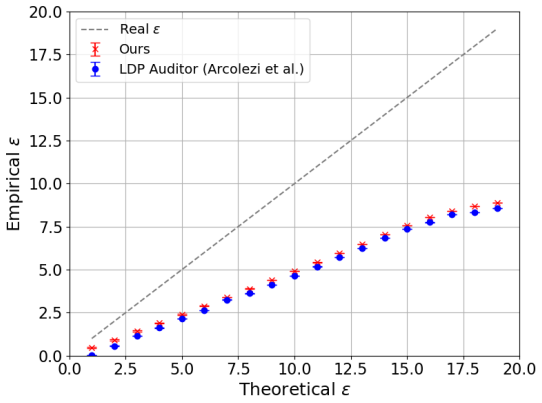
(a) Porto dataset.



(b) Beijing dataset.



(c) Porto dataset.



(d) Beijing dataset.

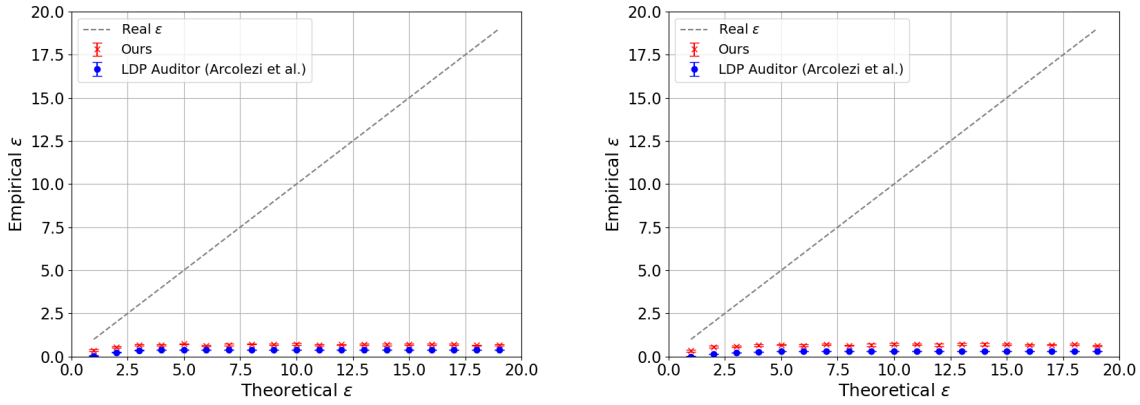
Figure 10.4.: Epsilon estimation for attacks on the HE mechanism: SHE (Attack 12) on the top row and THE (Attack 13) on the bottom row.

10.2. Identifying Bugs with U-ReRo

An important use case of DP auditing is the detection of implementation errors in DP algorithms [14]. To investigate whether U-ReRo-based auditing can reveal such violations, we first experiment with self-introduced bugs into our GRR implementation before turning to a real-world test-case.

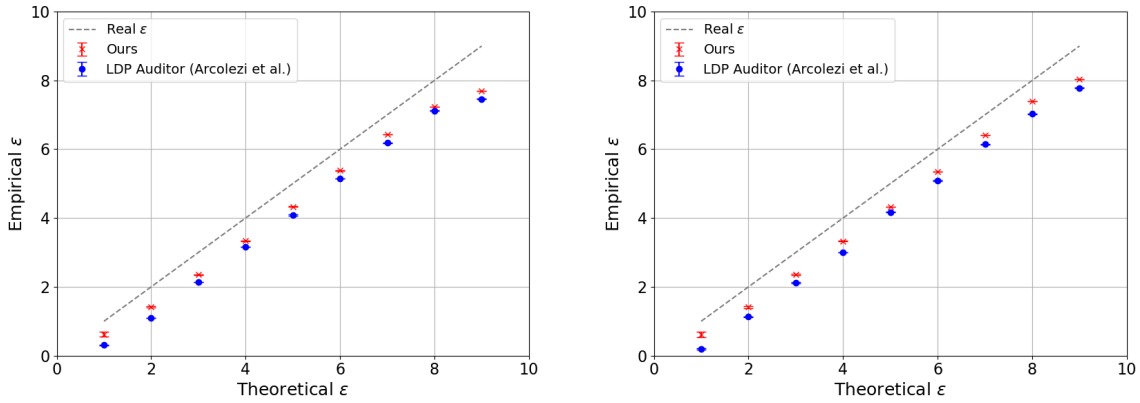
The results for our faulty implementation of GRR are plotted in Figure 10.6. The bug we introduced affects the $1 - p$ case: instead of sampling from the set of all nodes excluding the true node, the faulty implementation samples from the full graph, including the true node. For low values of the privacy budget $\varepsilon \in \{1, 2\}$, we can observe that the flawed implementation causes U-ReRo to exceed the theoretical bound. For the higher values, we cannot identify the flaw which is both due to the large domain sizes m and also because at higher values of ε, the probability p approaches 1, effectively eliminating the error case.

To amplify the impact of the error, we introduced a stronger violation: we artificially increased the probability p of reporting the true node by a constant factor of 0.1, thereby intentionally skewing the mechanism's behavior. The results are plotted in Figure 10.7. For



(a) Porto dataset.

(b) Beijing dataset.



(c) Porto dataset.

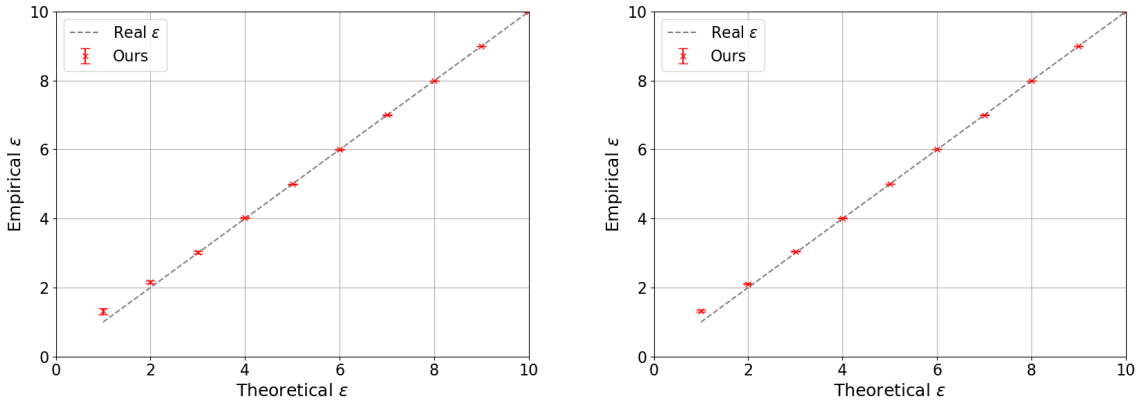
(d) Beijing dataset.

Figure 10.5.: Epsilon estimation for Attack 14 the LH mechanisms: BLH on the top row and OLH on the bottom row. OLH computed only until $\varepsilon = 10$ for computational feasibility.

the stronger violation, the estimation successfully identifies the flawed implementation: we observe consistent overestimation of the privacy budget for all values of $\varepsilon \leq 10$. U-ReRo thus allows us to clearly identify the implementation flaw in our GRR mechanism.

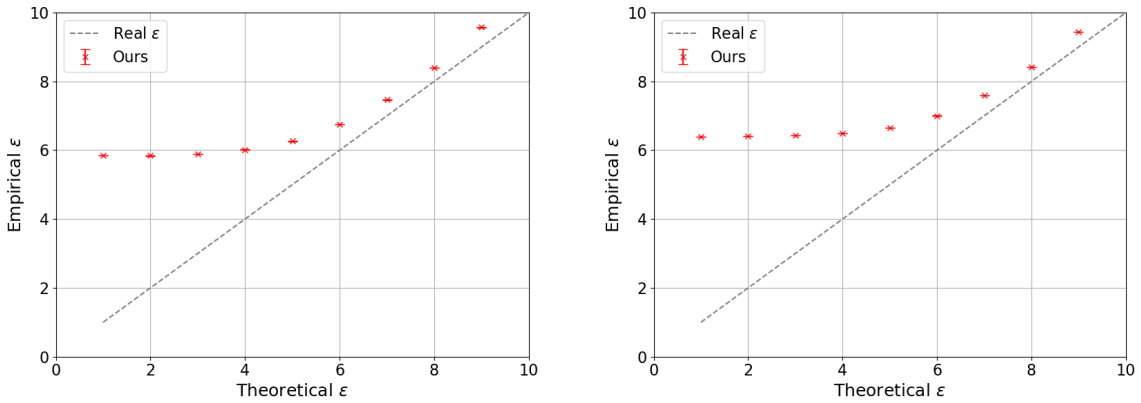
As a final step, we turn to a realistic setting to assess U-ReRo for bug detection. Specifically, we replicate the case study from [6], which demonstrates how their LDP Auditor uncovered an implementation bug in the pure-LDP Python package [59] (version 1.1.2), a widely used library that implements, among others, the UE mechanism. Arcolezi et al.'s evaluation of the SUE and OUE protocols via LDP Auditor revealed that the empirical privacy leakage exceeded the claimed guarantees for low privacy budget $\varepsilon \in \{0.25, 0.5\}$.

The bug in the flawed UE implementation resulted from a missing correction step: a user's vector is originally initialized as a zero-vector. First, each bit in the vector is flipped from 0 to 1 with probability q. Then, the user's true input value at position $j \in [m]$ is flipped from 0 to 1 with probability p. However, if the true bit at position j was already flipped in the first step but not again in the second, it should have been reset to 0 – this step was omitted. As a result, the true value was overrepresented, especially for low values



(a) Porto dataset. (b) Beijing dataset.

Figure 10.6.: Epsilon estimation for GRR attack with a subtle implementation flaw (sampling from the full graph including the true node in the $(1 - p)$ case).



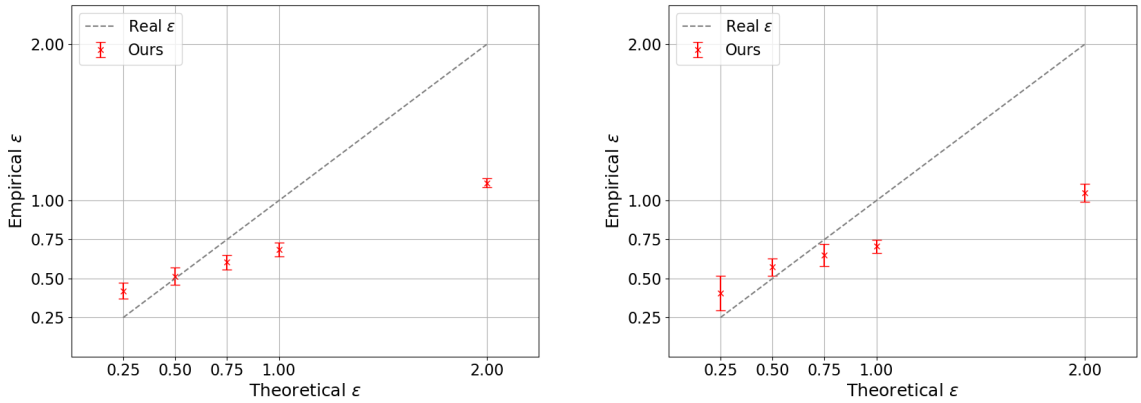
(a) Porto dataset. (b) Beijing dataset.

Figure 10.7.: Epsilon estimation for the GRR attack with a significantly flawed implementation (truthful reporting probability p artificially increased by 0.1).

of the privacy budgets where the probability of transmitting the user's true bit at position j as 1 is low [6].

In Figure 10.8 and Figure 10.9, we can see the auditing results for U-ReRo based on the flawed SUE and OUE implementations. We can see that across all experiments, we find an overestimation of the empirical privacy budget for $\varepsilon = 0.25$ and for all cases but SUE on the Porto dataset also for $\varepsilon = 0.5$, despite our data domains being larger than what was analyzed for LDP Auditor.

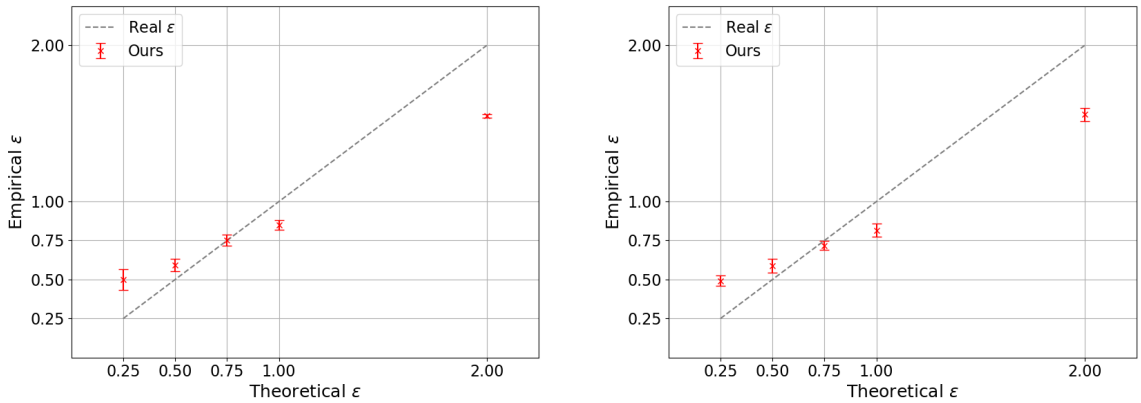
This result demonstrates that U-ReRo is sensitive enough to detect even subtle implementation flaws. The consistent overestimation of the empirical privacy budget in high-privacy settings ($\varepsilon = 0.25$), confirms that even small deviations from the intended protocol can measurably impact privacy guarantees. Importantly, this holds across both of our datasets and larger domains than those previously analyzed which highlights that U-ReRo can be used as a robust tool for DP auditing.



(a) Porto dataset.

(b) Beijing dataset.

Figure 10.8.: Epsilon estimation for the UE attack on flawed SUE implementation from pure-LDP package [59] (version 1.1.2).



(a) Porto dataset.

(b) Beijing dataset.

Figure 10.9.: Epsilon estimation for the UE attack on flawed OUE implementation from pure-LDP package [59] (version 1.1.2).

10.3. Intermediate Conclusions

We provided a general framework to use U-ReRo for DP auditing that allows to extend auditing beyond membership inference and indistinguishability attacks. We showed that U-ReRo-based auditing performs either comparably or even better than the state-of-the-art auditing tool for LDP, LDP Auditor, over sizeable data domains and a wide range of LDP protocols. U-ReRo-based auditing addresses two limitations of LDP Auditor: for one, while LDP Auditor audits based on indistinguishability attacks, our method extends DP auditing to any attack strategy. Secondly, U-ReRo-based auditing does not constrain the maximum estimatable value of the privacy budget due to its parameter choices, whereas LDP Auditor is limited by its choice of T and α. Finally, we demonstrated that U-ReRo is sufficiently precise as an auditing metric to detect even subtle implementation flaws in LDP mechanisms.

One limitation of U-ReRo-based auditing remains: it requires many repeated runs of the DP mechanism to provide a reliable estimate. While this is feasible in the context of lightweight LDP mechanisms, it poses a scalability challenge when applied to domains like private learning, where each run involves training a complete ML model. Recent work in the auditing literature has explored approaches that aim to perform auditing in a single or a limited number of runs [85, 61, 52]–an area that, while promising, lies beyond the scope of this work and is not directly compatible with U-ReRo's sampling-based design. Researching approaches that allow U-ReRo to be audited with fewer runs or adapt it to settings with higher computational costs would be a valuable next step for future work for U-ReRo-based DP auditing.

11. Conclusion

Implementations of DP in practice require selecting an appropriate privacy budget. However, this choice is not trivial, as practitioners often lack guidance on how a certain ε translates into real-world privacy guarantees. A key aspect of ε interpretability is evaluating how effectively it mitigates privacy attacks. For this, we rely on metrics to measure adversarial success. Previous metrics, such as the membership or attribute advantage, are limited to a certain type of attacks, however, ReRo and its extension U-ReRo work as general metrics and promise a unifying framework for attack resistance analysis.

In this work, we provided a thorough evaluation of ReRo and U-ReRo in the LDP and ML setting. Our empirical evaluation shows that ReRo actually overestimates privacy leakage by factoring in knowledge learned from population-level patterns or from background knowledge, rather than focusing on individual privacy leakage resulting from participation in the dataset. Such overestimation encourages overly conservative choices of the privacy budgets and sacrifices utility in the published data without a corresponding gain in privacy. We furthermore showed that U-ReRo successfully distinguishes individual leakage and hence provides a more accurate assessment of risk.

Importantly, the joint interpretation of ReRo and U-ReRo provides deeper insight. While ReRo reflects the total success probability of reconstruction–irrespective of its cause–U-ReRo quantifies the effect of individual participation. Together, they enable a more nuanced understanding of attack risk: one can infer how much of the threat stems from general correlations versus personal data exposure. As Jayaraman et al. argue, in some settings the origin of leakage may be irrelevant [47]–but when it is, U-ReRo helps distinguish and quantify individual risk.

We furthermore provided novel bounds on U-ReRo and showed that they improve over the state-of-the-art [37]. Our novel bounds are perfectly tight under uniform prior and also provide a good estimate of privacy leakage for non-uniform data distributions.

Moreover, we addressed the limitation of U-ReRo concerning *aux*– it does not incorporate target-specific auxiliary knowledge thus ignoring strong attackers in membership and attribute inference– with our novel metric Aux-Aware U-ReRo which incorporates target-specific auxiliary knowledge into the analysis of data reconstruction attacks. To the best of our knowledge, this is the first general reconstruction metric that naturally subsumes both membership and attribute inference as special cases. We established a bound on Aux-Aware U-ReRo and demonstrated that it is reasonably tight in practice. Exploring attacks that leverage *aux* to show the limitations of existing bounds on classic ReRo and U-ReRo in practice as well as deriving tighter bounds on Aux-Aware U-ReRo across diverse settings could be interesting future work.

Finally, we introduced a novel framework to employ U-ReRo for DP auditing. As a metric for data reconstruction attacks, U-ReRo allows auditing privacy leakage beyond membership inference, which dominates the current literature. Membership is not al-

ways sensitive and recent evidence suggests that forgoing protection against membership inference can still protect from broader attacks such as data reconstruction [10]. Our U-ReRo-based approach allows to extend DP auditing to any attack strategy and therefore addresses broader threat scenarios. We demonstrated that U-ReRo-based auditing performs on par with – and in some cases outperforms – the state-of-the-art auditing tool LDP Auditor across various attacks and mechanisms in the LDP setting under uniform prior. One limitation is that auditing via U-ReRo currently requires running the chosen attack many times and then evaluating the metric, making it less suited for scenarios where only a few attack runs are feasible. Addressing this limitation could be another important direction for future work.

Bibliography

[1] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 308–318. ISBN: 9781450341394.

[2] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. `https://www.tensorflow.org/`. 2015.

[3] John M. Abowd and Michael B. Hawes. "Confidentiality Protection in the 2020 US Census of Population and Housing". In: *Annual Review of Statistics and Its Application* 10.Volume 10, 2023 (2023), pp. 119–144.

[4] Miguel E. Andrés et al. "Geo-indistinguishability: differential privacy for location-based systems". In: *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. CCS '13. Berlin, Germany: Association for Computing Machinery, 2013, pp. 901–914. ISBN: 9781450324779.

[5] Meenatchi Sundaram Muthu Selva Annamalai et al. *The Hitchhiker's Guide to Efficient, End-to-End, and Tight DP Auditing*. 2025. arXiv: `2506.16666`.

[6] Héber H. Arcolezi and Sébastien Gambs. "Revealing the True Cost of Locally Differentially Private Protocols: An Auditing Perspective". In: *Proceedings on Privacy Enhancing Technologies* 2024.4 (Oct. 2024), pp. 123–141.

[7] Héber H. Arcolezi et al. "On the Risks of Collecting Multidimensional Data Under Local Differential Privacy". In: *Proceedings of the VLDB Endowment* 16.5 (Jan. 2023), pp. 1126–1139.

[8] A. Asuncion and D. J. Newman. *UCI Machine Learning Repository*. `http://www.ics.uci.edu/~mlearn/MLRepository.html`. 2007.

[9] Chiara Bachechi and Laura Po. "Road Network Graph Representation for Traffic Analysis and Routing". In: *Advances in Databases and Information Systems*. Ed. by Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel. Cham: Springer International Publishing, 2022, pp. 75–89. ISBN: 9783031157400.

[10] Borja Balle, Giovanni Cherubin, and Jamie Hayes. "Reconstructing Training Data with Informed Adversaries". In: *2022 IEEE Symposium on Security and Privacy (SP)*. 2022, pp. 1138–1156.

[11] Raef Bassily and Adam Smith. "Local, Private, Efficient Protocols for Succinct Histograms". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. STOC '15. Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 127–135. ISBN: 9781450335362.

[12] Daniel Bernau et al. "Comparing Local and Central Differential Privacy Using Membership Inference Attacks". In: *Data and Applications Security and Privacy XXXV*. Ed. by Ken Barker and Kambiz Ghazinour. Cham: Springer International Publishing, 2021, pp. 22–42. ISBN: 9783030812423.

[13] Daniel Bernau et al. *Quantifying identifiability to choose and audit ϵ in differentially private deep learning*. 2021. arXiv: 2103.02913.

[14] Benjamin Bichsel et al. "DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers". In: *2021 IEEE Symposium on Security and Privacy (SP)*. May 2021, pp. 391–409.

[15] Geoff Boeing. *Modeling and Analyzing Urban Networks and Amenities with OSMnx*. https://geoffboeing.com/publications/osmnx-paper/. 2024.

[16] Mark Bun and Thomas Steinke. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: *Theory of Cryptography*. Ed. by Martin Hirt and Adam Smith. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 635–658. ISBN: 9783662536414.

[17] Mark Bun et al. *Statistical inference is not a privacy violation*. https://differentialprivacy.org/inference-is-not-a-privacy-violation/. 2021.

[18] Nicholas Carlini et al. "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 2633–2650. ISBN: 9781939133243.

[19] Rachel Cummings et al. "ATTAXONOMY: Unpacking Differential Privacy Guarantees Against Practical Adversaries". In: *ArXiv* (2024). arXiv: 2405.01716.

[20] Edoardo Debenedetti et al. "Privacy Side Channels in Machine Learning Systems". In: *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 6861–6848. ISBN: 9781939133441.

[21] E. W. Dijkstra. "A note on two problems in connexion with graphs". In: *Numer. Math.* 1.1 (Dec. 1959), pp. 269–271.

[22] Youlong Ding et al. *Delving into Differentially Private Transformer*. 2024. arXiv: 2405.18194.

[23] Zeyu Ding et al. "Detecting Violations of Differential Privacy". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS '18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 475–489. ISBN: 9781450356930.

[24] Jinshuo Dong, Aaron Roth, and Weijie J. Su. "Gaussian Differential Privacy". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.1 (Feb. 2022), pp. 3–37.

[25] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. "Local Privacy and Statistical Minimax Rates". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. 2013, pp. 429–438.

[26] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. “Minimax Optimal Procedures for Locally Private Estimation”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 182–201.

[27] Cynthia Dwork. “Differential Privacy”. In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 9783540359081.

[28] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy”. In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[29] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[30] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”. In: CCS ’14. Scottsdale, Arizona, USA: Association for Computing Machinery, 2014, pp. 1054–1067. ISBN: 9781450329576.

[31] Úlfar Erlingsson et al. “That which we call private”. In: *ArXiv* (2019). arXiv: 1908.03566.

[32] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “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. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. ISBN: 9781450338325.

[33] Matthew Fredrikson et al. “Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing”. In: *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 17–32. ISBN: 9781931971157.

[34] Andrea Gadotti et al. “Pool Inference Attacks on Local Differential Privacy: Quantifying the Privacy Guarantees of Apple’s Count Mean Sketch in Practice”. In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 501–518. ISBN: 9781939133311.

[35] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. “Show me how you move and I will tell you who you are”. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. SPRINGL ’10. San Jose, California: Association for Computing Machinery, 2010, pp. 34–41. ISBN: 9781450304351.

[36] Elena Ghazi and Ibrahim Issa. “Total Variation Meets Differential Privacy”. In: *IEEE Journal on Selected Areas in Information Theory* 5 (2024), pp. 207–220.

[37] Patricia Guerra-Balboa, Annika Sauer, and Thorsten Strufe. "Analysis and Measurement of Attack Resilience of Differential Privacy". In: *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*. WPES '24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 155–171. ISBN: 9798400712395.

[38] Chuan Guo et al. "Bounding Training Data Reconstruction in Private (Deep) Learning". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 8056–8071.

[39] M. Emre Gursoy et al. "An Adversarial Approach to Protocol Analysis and Selection in Local Differential Privacy". In: *IEEE Transactions on Information Forensics and Security* 17 (2022), pp. 1785–1799.

[40] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. "Exploring network structure, dynamics, and function using NetworkX". In: Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). Jan. 2008.

[41] Richard W. Hamming. *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[42] Jamie Hayes, Borja Balle, and Saeed Mahloujifar. "Bounding training data reconstruction in DP-SGD". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 78696–78722.

[43] Florimond Houssiau et al. *TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data*. 2022. arXiv: 2211.06550.

[44] Thomas Humphries et al. "Investigating Membership Inference Attacks under Data Dependencies". In: *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. 2023, pp. 473–488.

[45] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. "Auditing differentially private machine learning: How private is private sgd?" In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22205–22216.

[46] Bargav Jayaraman. "Analyzing the Leaky Cauldron: Inference Attacks on Machine Learning". Available at: https://libraetd.lib.virginia.edu/public$_{v}iew/1r66j21378$. Ph.D. dissertation. University of Virginia, Dec. 2022.

[47] Bargav Jayaraman and David Evans. "Are Attribute Inference Attacks Just Imputation?" In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. CCS '22. Los Angeles, CA, USA: Association for Computing Machinery, 2022, pp. 1569–1582. ISBN: 9781450394505.

[48] Bargav Jayaraman and David Evans. "Evaluating Differentially Private Machine Learning in Practice". In: *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1895–1912. ISBN: 9781939133069.

[49] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. “Discrete distribution estimation under local privacy”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML’16. New York, NY, USA: JMLR.org, 2016, pp. 2436–2444.

[50] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. “The Composition Theorem for Differential Privacy”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1376–1385.

[51] Shiva Prasad Kasiviswanathan et al. “What Can We Learn Privately?” In: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826.

[52] Amit Keinan, Moshe Shenfeld, and Katrina Ligett. *How Well Can Differential Privacy Be Audited in One Run?* 2025. arXiv: 2503.07199.

[53] Jong Wook Kim, Dae-Ho Kim, and Beakcheol Jang. “Application of Local Differential Privacy to Collection of Indoor Positioning Data”. In: *IEEE Access* 6 (2018), pp. 4276–4286.

[54] Dirk P. Kroese et al. “Why the Monte Carlo method is so important today”. In: *WIREs Computational Statistics* 6.6 (2014), pp. 386–392.

[55] John Krumm and Dany Rouhana. “Placer: semantic place labels from diary data”. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’13. Zurich, Switzerland: Association for Computing Machinery, 2013, pp. 163–172. ISBN: 9781450317702.

[56] Bogdan Kulynych et al. *Unifying Re-Identification, Attribute Inference, and Data Reconstruction Risks in Differential Privacy*. 2025. arXiv: 2507.06969.

[57] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

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

[59] Samuel Maddock. *pure-LDP*. https://pypi.org/project/pure-ldp/. 2021.

[60] Samuel Maddock, Alexandre Sablayrolles, and Pierre Stock. “CANIFE: Crafting Canaries for Empirical Privacy Measurement in Federated Learning”. In: *ArXiv* (2022). arXiv: 2210.02912.

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

[62] Saeed Mahloujifar et al. *Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms*. 2022. arXiv: 2204.06106.

[63] Mani Malek Esmaeili et al. “Antipodes of Label Differential Privacy: PATE and ALIBI”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 6934–6945.

[64] Frank McSherry. *Statistical inference considered harmful*. https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md. 2016.

[65] Frank McSherry and Kunal Talwar. "Mechanism Design via Differential Privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. 2007, pp. 94–103.

[66] Shagufta Mehnaz et al. "Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 4579–4596. ISBN: 9781939133311.

[67] Àlex Miranda-Pascual et al. "SoK: Differentially private publication of trajectory data". In: *Proceedings on Privacy Enhancing Technologies* (2023).

[68] Yves-Alexandre de Montjoye et al. "Unique in the Crowd: The privacy bounds of human mobility". In: *Scientific Reports* 3 (2013).

[69] Meenatchi Sundaram Muthu Selva Annamalai. "It's Our Loss: No Privacy Amplification for Hidden State DP-SGD With Non-Convex Loss". In: *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*. AISec '24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 24–30. ISBN: 9798400712289.

[70] Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. "Nearly Tight Black-Box Auditing of Differentially Private Machine Learning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 131482–131502.

[71] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 111–125.

[72] Milad Nasr et al. "Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning". In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021, pp. 866–882.

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

[74] Keiron O'Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. 2015. arXiv: `1511.08458`.

[75] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. `https://www.openstreetmap.org`. 2017.

[76] Art B. Owen. *Monte Carlo theory, methods and examples*. `https://artowen.su.domains/mc/`, 2013.

[77] Nicolas Papernot et al. *Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data*. 2017. arXiv: `1610.05755`.

[78] Natalia Ponomareva et al. "How to DP-fy ML: A Practical Tutorial to Machine Learning with Differential Privacy". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. Long Beach, CA, USA: Association for Computing Machinery, 2023, pp. 5823–5824. ISBN: 9798400701030.

[79] IUrıı Vasilevich Prokhorov and V Statulevicius. *Limit theorems of probability theory*. Vol. 6. Springer Science & Business Media, 2000.

[80] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. "What Does The Crowd Say About You? Evaluating Aggregation-based Location Privacy". In: *Proceedings on Privacy Enhancing Technologies* 2017 (2017), pp. 156–176.

[81] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[82] Donald B Rubin. "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse". In: *Proceedings of the survey research methods section of the American Statistical Association*. Vol. 1. American Statistical Association Alexandria, VA. 1978, pp. 20–34.

[83] Donald B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (Dec. 1976), pp. 581–592.

[84] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18.

[85] Thomas Steinke, Milad Nasr, and Matthew Jagielski. "Privacy auditing with one (1) training run". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2023.

[86] Latanya Sweeney. *Simple Demographics Often Identify People Uniquely*. Data Privacy Working Paper 3. Carnegie Mellon University, Data Privacy Lab, 2000.

[87] Shun Takagi et al. "Geo-Graph-Indistinguishability: Protecting Location Privacy for LBS over Road Networks". In: *Data and Applications Security and Privacy XXXIII*. Ed. by Simon N. Foley. Cham: Springer International Publishing, 2019, pp. 143–163. ISBN: 9783030224790.

[88] Florian Tramèr et al. "Debugging Differential Privacy: A Case Study for Privacy Auditing". In: *ArXiv* (2022). arXiv: 2202.12219.

[89] Jincheng Wang et al. "Topology-theoretic approach to address attribute linkage attacks in differential privacy". In: *Computers Security* 113 (2022), p. 102552.

[90] Tianhao Wang et al. "Locally differentially private protocols for frequency estimation". In: *26th USENIX Security Symposium (USENIX Security 17)*. 2017, pp. 729–745.

[91] Stanley L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69.

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

[93] Yonghui Xiao and Li Xiong. "Protecting Locations with Differential Privacy under Temporal Correlations". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1298–1309. ISBN: 9781450338325.

[94] Yonghui Xiao, Li Xiong, and Chun Yuan. "Differentially Private Data Release through Multidimensional Partitioning". In: *Secure Data Management*. Ed. by Willem Jonker and Milan Petković. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 150–168. ISBN: 9783642155468.

[95] Fengli Xu et al. "Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[96] Min Ye and Alexander Barg. "Optimal Schemes for Discrete Distribution Estimation Under Locally Differential Privacy". In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5662–5676.

[97] Samuel Yeom et al. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)* (2017), pp. 268–282.

[98] Benjamin Zi Hao Zhao et al. "On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models". In: *2021 IEEE European Symposium on Security and Privacy (EuroSP)*. 2021, pp. 232–251.

[99] Xiangguo Zhao et al. "LDPart: Effective Location-Record Data Publication via Local Differential Privacy". In: *IEEE Access* 7 (2019), pp. 31435–31445.

[100] Yu Zheng et al. *Geolife GPS trajectory dataset 1.1 - User Guide*. `https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/`. 2011.