Enhancing Local Differential Privacy Accuracy by Exploiting Inherent Uncertainty

PENG TANG*, School of Cyber Science and Technology, Shandong University, China
XIYA SHAO*, School of Cyber Science and Technology, Shandong University, China
RUI CHEN, College of Computer Science and Technology, Harbin Engineering University, China
NING WANG, Cyberspace Institute of Advanced Technology, Guangzhou University, China
SHANQING GUO*†, School of Cyber Science and Technology, Shandong University, China

Local differential privacy (LDP) is a standard for privacy preservation in the local setting: classical LDP mechanisms protect directly sensitive attributes by injecting uncertainty between users' true values and reported data. In many deployments, however, practitioners also choose to perturb attributes that are not inherently sensitive but are correlated with latent sensitive ones, using LDP as a system-level choice to limit attribute-inference risks. When such correlated attributes are naively treated as fully sensitive, the resulting mechanisms may satisfy a given privacy requirement but inject more noise than necessary, causing utility loss. This raises the challenge of quantifying how much uncertainty correlations already provide about sensitive attributes and using this to optimize perturbation under fixed privacy constraints.

We address this challenge via *inherent uncertainty*, a metric that captures correlation-induced uncertainty between collected non-sensitive attributes and associated sensitive attributes. We also introduce a correlation-aware LDP framework that exploits this metric to recalibrate perturbation parameters, while providing the same ε-LDP guarantee for sensitive attributes as a treat-as-sensitive baseline. We further consider a hybrid scenario in which a single collected attribute contains both sensitive values and non-sensitive-but-correlated values, and propose a dual-phase perturbation method that provides differentiated protection and improves utility. Our mechanisms satisfy ε-LDP and the specified inference guarantees, and experiments on real datasets demonstrate their utility gains. Code is available at https://github.com/null944/enhanced-accuracy-in-ldp.

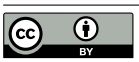
CCS Concepts: • **Security and privacy → Data anonymization and sanitization**.

Additional Key Words and Phrases: Local differential privacy; Data collection; Inherent uncertainty

ACM Reference Format:
Peng Tang, Xiya Shao, Rui Chen, Ning Wang, and Shanqing Guo. 2026. Enhancing Local Differential Privacy Accuracy by Exploiting Inherent Uncertainty. *Proc. ACM Manag. Data* 4, 1 (SIGMOD), Article 33 (February 2026), 26 pages. https://doi.org/10.1145/3786647

*Also with State Key Laboratory of Cryptography and Digital Economy Security, Shandong University, Qingdao, China, and Key Laboratory of Cryptologic Technology and Information Security of Ministry of Education, Shandong University, Qingdao, China.
†Corresponding author.

Authors' Contact Information: Peng Tang, School of Cyber Science and Technology, Shandong University, Qingdao, China, tangpeng@sdu.edu.cn; Xiya Shao, School of Cyber Science and Technology, Shandong University, Qingdao, China, shaoxiya@mail.sdu.edu.cn; Rui Chen, College of Computer Science and Technology, Harbin Engineering University, Harbin, China, ruichen@hrbeu.edu.cn; Ning Wang, Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China, wangning@gzhu.edu.cn; Shanqing Guo, School of Cyber Science and Technology, Shandong University, Qingdao, China, guoshanging@sdu.edu.cn.



ACM 2836-6573/2026/2-ART33
https://doi.org/10.1145/3786647

1 Introduction

Differential privacy (DP) [11, 12] has increasingly become the standard for data privacy. Recently, techniques for satisfying DP in the local setting, referred to as local differential privacy (LDP), have been developed. LDP facilitates the collection of statistics from decentralized users while protecting privacy without necessitating a trusted third party.

Classical LDP mechanisms target attributes that are *directly sensitive*, such as income or diagnosis: a user perturbs a sensitive attribute S so that the reports satisfy ε-LDP with respect to S. This direct-disclosure setting is the standard LDP use case studied in the DP/LDP literature and provides the reference setting for our work. In many real-world deployments, curators collect a non-sensitive attribute A that is statistically correlated with a latent sensitive attribute S. Releasing such an attribute A *without additional protection* can enable attackers to infer S by exploiting these correlations, for example inferring mental health from sleep patterns or home and workplace locations from mobility traces. We term this correlation-induced inference threat a *deployment-level indirect inference risk* about the latent sensitive attribute S and, for brevity, refer to it as an *indirect privacy risk*.

From a DP standpoint, these indirect risks stem from data correlations, not from any flaw in LDP. In practice, system designers often choose to apply LDP to such an attribute A to limit inference about S, using it as a building block beyond what the standard definition of LDP is formally required to guarantee. A straightforward mitigation is to treat A as fully sensitive and apply a standard ε-LDP protocol to generate a perturbed output $\tilde{A}$. While this baseline can be tuned to bound inference risk on the sensitive attribute S, the "treat-as-sensitive" strategy is often over-conservative: it disregards the inherent uncertainty in the correlation between A and S, and thus injects more noise than necessary, causing avoidable utility loss. This motivates our central question: *given a fixed privacy requirement that the overall process of publishing $\tilde{A}$ must satisfy an ε-LDP requirement for the latent sensitive attribute S, can we design LDP mechanisms that achieve the same privacy guarantee with less perturbation and higher utility than the treat-as-sensitive baseline?*

Our key idea is that the correlation between A and S already imposes a form of *inherent uncertainty* on any inference of S, while LDP injects additional randomness between true values and reports. We formalize this by introducing **Inherent Uncertainty**, an information-theoretic metric that quantifies how much uncertainty about S remains given A. We then develop a correlation-aware LDP framework that *recalibrates* the perturbation parameters of standard LDP mechanisms: for this fixed privacy requirement, our mechanisms provably provide the same ε-LDP guarantee for the sensitive attribute S as the baseline design that treats A as fully sensitive, while using significantly less noise. In this sense, we operate strictly within the LDP framework and focus on *utility optimization under fixed privacy*, rather than modifying or weakening the underlying notion of privacy. We instantiate this framework with concrete LDP protocols that tailor perturbation to the actual inference threat rather than to a worst-case assumption.

Furthermore, we address a **hybrid scenario** in which the collected attribute A takes two kinds of values in its domain: sensitive values and non-sensitive values that are statistically correlated with a latent sensitive attribute S. For example, when collecting *cholesterol level*, the measured value is sensitive, while its missing status is non-sensitive in isolation but may be correlated with another sensitive attribute S such as *age*. In such cases, a single uniform perturbation strategy is inadequate, because it must simultaneously control the direct privacy risk from sensitive values of A and the indirect privacy risk carried by the missing status of A about S. To handle such leakage-pattern-driven privacy problems, we propose a dual-phase perturbation framework. It first injects an "equivalent" amount of uncertainty into sensitive values of A, derived from the inherent uncertainty of correlated non-sensitive values, and then applies a unified LDP perturbation

to all values, providing differentiated protection while still satisfying the same specified privacy requirements on sensitive values of A and on the latent attribute S as the corresponding treat-as-sensitive baselines.

Our key contributions are summarized as follows.

- We formalize indirect and hybrid *deployment-level* privacy risks when collecting correlated attributes, and cast these deployment scenarios as utility-optimization problems under fixed privacy requirements, strictly within the DP/LDP framework.
- We introduce a correlation-aware LDP framework. This framework exploits Inherent Uncertainty to recalibrate perturbation parameters, providing the same ε-LDP guarantee for the sensitive attribute S as the treat-as-sensitive baseline (which applies standard LDP to attribute A as if it were fully sensitive), while substantially improving utility.
- We propose a dual-phase perturbation scheme for hybrid scenarios where different values of the same attribute induce heterogeneous privacy risks. This scheme provides a principled way to offer differentiated protection under the same privacy constraints while maximizing utility.
- We present privacy and utility analyses of the proposed schemes and empirically validate their effectiveness through extensive experiments on real datasets.

2 Related work

2.1 Direct Privacy Leakage Risks

Current LDP protocols primarily address the direct leakage of sensitive data. The data collected with LDP can be categorical or numerical/ordinal. Categorical data can be classified into binary attribute data [35] and multivalued attribute data [3, 13, 18, 19] based on an attribute's domain size. For numerical data, some mechanisms [22, 31] consider distribution estimation under LDP, while some others mainly focus on mean estimation [6, 10, 30]. Some works [5, 38] exist where each user has multiple attributes, and the aggregator is interested in the joint distribution of some attributes. In practical applications, a trade-off exists between privacy protection and the accuracy of aggregated results in LDP. To address this challenge, various approaches have been proposed, including parameter optimization [32], sampling methods [21, 27], post-processing methods [9, 34], shuffling frameworks [4], and popular division techniques [28, 30], among others.

2.2 Indirect Privacy Leakage Risks

An *indirect privacy leakage risk* arises in deployment settings where a non-sensitive collected attribute A is statistically correlated with a latent sensitive attribute S.

Information theory-based approach. There are some existing works [8, 16, 17, 37] that address this problem using information-theoretic techniques. In particular, Calmon et al. [8] establish a general statistical inference framework, proposing two privacy metrics: average information leakage and maximum information leakage. Jiang et al. [17] introduce the local information privacy (LIP) model, which quantifies privacy via the divergence between prior and posterior probabilities of sensitive attributes. Hsu et al. [16] employ *lift* to measure privacy risks introduced by correlated attributes, identifying users who can safely disclose their data. Zarrabian et al. [37] discover lift asymmetry, leading to the Asymmetric Local Information Privacy (ALIP) model, which sets distinct privacy bounds for max/min-lift. *These information-theoretic metrics provide a foundational language for quantifying privacy uncertainty. Our work on "inherent uncertainty" is philosophically aligned with these concepts. However, our core contribution diverges in its goal and output. Prior works in this space often aim to define new privacy definitions or provide alternative semantic guarantees. Our work, in contrast, introduces a novel mechanism design within the established LDP framework. We translate the information-theoretic insight into a concrete parameter that serves as a pragmatic design knob for*

configuring perturbation. Consequently, while inspired by information theory, our approach outputs standard ε-LDP guarantees, yet achieves a superior privacy-utility trade-off by being explicitly aware of correlation-based risks.

Differential privacy-based approach. Building on differential privacy, several works [23, 24] primarily focus on deriving global information-theoretic formulas that quantify inference risks. *In contrast, our work not only refines this quantification perspective, but also develops concrete optimized LDP protocols and extends the analysis to a leakage-pattern-driven hybrid privacy setting*, where the attribute domain contains both sensitive values and non-sensitive values that remain statistically dependent on other sensitive attributes. For such a hybrid scenario, we emphasize the design of perturbation mechanisms that satisfy heterogeneous privacy requirements across different values. Existing works on personalized local differential privacy [1, 2, 14, 15, 26] also allow attribute-specific privacy guarantees. *However, in those works heterogeneous requirements are driven by differences in value sensitivity, whereas in our setting they are induced by distinct privacy leakage patterns.*

2.3 Collection of Multiple Correlated Sensitive Attributes

For the collection of multiple related sensitive attributes, prior works [7, 36] study scenarios where statistical correlations exist among sensitive attributes. When an attacker observes a subset of attributes, they may infer others that have not yet been collected; these works typically treat such "information inferable from known attributes" as prior knowledge that is inherently non-private and therefore not explicitly protected. *Our setting is complementary: we focus on collecting just a single non-sensitive attribute that is correlated with sensitive ones, and design protection mechanisms that explicitly aim to block implicit inference chains from this attribute to latent sensitive attributes.*

2.4 Generation/Collection of Missing Data

There are also works on missing data generation and collection under differential privacy. In the centralized setting, Mohapatra *et al.* [25] model missingness as a sampling process to obtain tighter privacy bounds for generating missing values from the ground-truth data. *By contrast, our goal is to collect missing data in the local setting.* For local mean estimation with missing data, Sun *et al.* [29] propose a bi-directional sampling mechanism for perturbation. *Our work considers two distinct types of privacy risks associated with missing and non-missing values and designs a tiered perturbation mechanism to address both, demonstrating its advantages through experimental comparisons.*

3 Preliminaries

3.1 Local Differential Privacy

Local differential privacy (LDP), the variant of differential privacy [11, 12], addresses the issue that the aggregator may not be trustworthy by considering a scenario in which each user i has only a local view of the dataset D, specifically, their own data point v_i. In order to protect privacy, each user i distorts their input value v_i using a function $\psi(\cdot) : v \to \widetilde{v}$, where v and $\widetilde{v}$ are the domains of the input and output values, and releases $\widetilde{v_i} = \psi(v_i)$ to the aggregator. This approach ensures that the individual data points remain private, even if the aggregator is malicious.

Definition 3.1 (ε-Local Differential Privacy [20]). Let v be the input domain, representing all possible data values of a single user. A randomized function $\psi : v \to \widetilde{v}$ satisfies ε-local differential privacy (LDP), where $\varepsilon \geq 0$, if and only if (iff) for any pair of neighboring input values $v_1, v_2 \in v$ (i.e., any two possible values in the input domain), and for all possible output subsets $O \subseteq \widetilde{v}$, the following inequality holds:

$$\Pr\left(\psi(v_1) \in O\right) \leq e^{\varepsilon} \cdot \Pr\left(\psi(v_2) \in O\right).$$

3.2 Pure LDP Protocols

A pure LDP protocol [33] requires the formal specification of a Support function that establishes the correspondence between each possible output $\widetilde{v} \in \widetilde{\nu}$ and the input set that $\widetilde{v}$ supports, referred to as $supp\,(\widetilde{v}) = \{v \in \nu \mid \widetilde{v} \text{ supports } v\}$.

Definition 3.2. (Pure LDP Protocols [33]). A protocol given by ψ and Support is pure iff there exist two probability values $p^* > q^*$ such that for all v_1,

$$\Pr\left(\psi\left(v_1\right) \in \{\widetilde{v} \mid v_1 \in supp\,(\widetilde{v})\}\right) = p^*,$$
$$\forall_{v_2 \neq v_1} \Pr\left(\psi\left(v_2\right) \in \{\widetilde{v} \mid v_1 \in supp\,(\widetilde{v})\}\right) = q^*.$$

By definition, for any given $\tilde{v} \in \widetilde{\nu}$, and any $v_1, v_2 \in \nu$, $\Pr\left(\tilde{v} \mid v_1\right) = \Pr\left(\tilde{v} \mid v_2\right)$ iff

$$(v_1 \in supp(\tilde{v}) \wedge v_2 \in supp(\tilde{v})) \vee (v_1 \notin supp(\tilde{v}) \wedge v_2 \notin supp(\tilde{v})).$$

We denote this by

$$\Pr\left(\tilde{v} \mid v\right) = \begin{cases} p^\circ, & \text{if } v \in supp\left(\tilde{v}\right), \\ q^\circ, & \text{if } v \notin supp\left(\tilde{v}\right), \end{cases} \tag{1}$$

and $p^\circ > q^\circ$. In the following, we will review some state-of-the-art pure LDP protocols.

3.2.1 *Randomized Response.*

Randomized Response (RR) [35] is a classic technique for implementing LDP where the input domain size is 2, i.e., $\nu = \{0, 1\}$.

- Encode and perturb. The input value v is still encoded as v, i.e., $\mathcal{C}\left(v\right) = v$ and the perturbation $\mathcal{P}$ is defined as:

$$\Pr\left(\mathcal{P}\left(v\right) = \widetilde{v}\right) = \begin{cases} p = \frac{e^\varepsilon}{1+e^\varepsilon} & \text{if } \widetilde{v} = v \\ q = \frac{1}{1+e^\varepsilon} & \text{if } \widetilde{v} \neq v \end{cases}. \tag{2}$$

- Aggregate. The aggregator takes all the reported values and obtains the count of $i \in \{0, 1\}$, denoted as c_i, and the estimate of the true count of i is $\widetilde{c_i} = \frac{c_i - nq}{p-q}$, where n denotes the number of users.

  In particular, for RR, the support sets are given by

$$supp\,(\widetilde{v} = 0) = \{0\} \text{ and } supp\,(\widetilde{v} = 1) = \{1\}.$$

3.2.2 *Generalized Randomized Response.*

The Generalized Randomized Response (GRR) [35] protocol extends RR to the case where the input domain size is more than 2 while satisfying ε-LDP. The input domain is referred to as $[d] = \{1, 2, \ldots, d\}$. Given a value v, GRR outputs the true value with probability p, and the probability for each remaining value is q. More formally, the perturbation function is defined as:

$$\Pr\left(\mathcal{P}\left(v\right) = \widetilde{v}\right) = \begin{cases} p = \frac{e^\varepsilon}{d-1+e^\varepsilon} & \text{if } \widetilde{v} = v \\ q = \frac{1}{d-1+e^\varepsilon} & \text{if } \widetilde{v} \neq v \end{cases}. \tag{3}$$

In particular, for GRR, $\forall i \in [d],\ supp\,(\widetilde{v} = i) = \{i\}$.

3.2.3 *Unary Encoding.*

In unary encoding (UE), a value is encoded as a bit vector:

- Encode. The value v is encoded as $B = [0, \ldots, 0, 1, 0, \ldots, 0]$, a length-$d$ binary vector where only the v-th position is 1.

- Perturb. $\mathcal{P}(B)$ outputs $\widetilde{B}$ as follows:

$$\Pr\left(\widetilde{B}[i]=1\right)=\begin{cases}p, & \text{if } B[i]=1\\ q, & \text{if } B[i]=0\end{cases}, \tag{4}$$

  To guarantee ε–LDP, we have $\frac{p}{q}\cdot\frac{1-q}{1-p}=e^{\varepsilon}$. To minimize the error of estimation, we can set $p=\frac{1}{2}$ and $q=\frac{1}{1+e^{\varepsilon}}$. Then, UE is referred to as Optimized UE (OUE) [33].
- Aggregate. The aggregator takes all the reported bit vectors and obtains the count of $\widetilde{B}$ whose $\widetilde{B}[i]=1$, denoted as c_i, and the estimate of the true count of i is $\widetilde{c}_i=\frac{c_i-nq}{p-q}$.

In particular, for OUE, we have $supp\left(\widetilde{B}\right)=\left\{i \mid \widetilde{B}[i]=1\right\}$.

4 Indirect Privacy Risk Prevention

The collection of correlated non-sensitive attributes can lead to indirect privacy risk. In this section, we first formally define this problem. Next, we propose an enhanced pure LDP protocol framework that leverages *inherent uncertainty*. Finally, we elaborate on how to quantify *inherent uncertainty* and recalculate perturbation parameters within this framework by integrating concrete LDP protocols.

4.1 Motivating Examples

Consider a scenario where a data collector gathers users' *exercise frequency* (non-sensitive attribute A). While A is non-sensitive, it is correlated with an *uncollected* sensitive attribute S, such as *diabetes status*. An adversary, upon receiving a user's exercise data A released without perturbation, can exploit the known statistical correlation to infer the user's sensitive diabetes status S. This exemplifies *indirect privacy risk*, the risk of inferring an uncollected sensitive attribute from a collected, correlated non-sensitive one.

4.2 Problem Definition and Analysis

Problem 1. *There are n users, each holding one data point that contains some attributes. Among these attributes, two are denoted as S and A, with domains Ω_S and Ω_A, respectively. Specifically, S is a sensitive attribute, and A is a correlated non-sensitive attribute. This means that disclosing the value of A will not directly reveal sensitive information but may indirectly leak information about S. An aggregator wants to collect frequency of A from users' data while satisfying ε-LDP for S, i.e., for any perturbed value $\widetilde{a}\in\Omega_{\widetilde{A}}$ and $s_i, s_j\in\Omega_S$, satisfies*

$$\Pr\left(\widetilde{A}=\widetilde{a}|S=s_i\right)\le e^{\varepsilon}\cdot\Pr\left(\widetilde{A}=\widetilde{a}|S=s_j\right). \tag{5}$$

Privacy model. Our work operates under the following privacy model and threat assumptions:

- Data Collection Scenario: An untrusted data curator aims to collect the *non-sensitive* attribute A from a population of users. We assume that A is statistically correlated with a *sensitive* attribute S, which is neither collected nor uploaded by any user. The goal of the collection is to obtain accurate aggregate statistics about A.
- Local Perturbation: Following the standard Local Differential Privacy (LDP) framework, each user locally perturbs their value a of attribute A using a randomized perturbation mechanism $\mathcal{P}$. The perturbed value $\tilde{a}=\mathcal{P}(a)$ is then sent to the curator. The raw value a is never disclosed.
- Adversarial Capabilities and Goals: The primary adversary is the untrusted curator. We also consider any external attacker who, like the curator, gains access to the perturbed $\tilde{a}$ and possesses the public knowledge of the conditional distribution $\Pr(A|S)$ (e.g., from a public dataset). The

goal of these adversaries is to infer an individual user's sensitive attribute S from $\tilde{a}$, exploiting the known A-S correlation.

- Privacy Goal: In this deployment, the system designer fixes a privacy requirement that the overall process of publishing $\tilde{A}$ must provide ε-LDP for the sensitive attribute S. A natural baseline is to treat A as fully sensitive and apply a standard ε-LDP mechanism to A, choosing its parameters so that this requirement is satisfied. This baseline is privacy-sound but can be conservative for utility, because it does not take into account the inherent uncertainty introduced by the correlation between A and S and may therefore add more noise than necessary. Our goal is to choose perturbation parameters for $\mathcal{P}$ that still ensure ε-LDP for S while improving the utility of the collected attribute A.

Analysis. In LDP, users perturb their inputs to introduce uncertainty and prevent attackers from inferring users' privacy based on the outputs. In the problem of correlated non-sensitive attribute collection, there exist three distinct information categories: (1) sensitive attribute S (attack targets), (2) correlated non-sensitive attribute A (protocol inputs), and (3) observed attribute $\widetilde{A}$ (protocol outputs). Two types of uncertainty exist between these categories:

(1) **Inherent uncertainty:** Inherent uncertainty, a concept inspired by (but operationally distinct from) information-theoretic measures of privacy, refers to the uncertainty that naturally exists within the data. This includes, for example, the uncertainty between a sensitive attribute S and a correlated non-sensitive attribute A.
(2) **Perturbation uncertainty:** Perturbation uncertainty is the uncertainty introduced through LDP.

The definition of our metric, **Inherent Uncertainty**, assumes knowledge of the conditional distribution $\Pr(A|S)$. This distribution can be obtained through several approaches: (1) Leveraging publicly available data to estimate $\Pr(A|S)$; (2) Applying incremental learning techniques, where an initial data subset is used to approximate the distribution. It quantifies the resulting privacy buffer, enabling our distribution-aware model for utility optimization.

The combined effect of *inherent uncertainty* and *perturbation uncertainty* provides privacy protection for the latent sensitive attribute S. To minimize perturbation uncertainty while maintaining ε-LDP, it is crucial to accurately quantify inherent uncertainty and establish a unified measurement framework for both uncertainty types.

4.3 Enhanced Pure LDP Protocols

Building upon the above analysis, we propose an enhanced pure LDP protocol framework for the problem of correlated non-sensitive attribute collection that fundamentally differs from existing protocols in two key aspects: (1) our framework initiates with precise quantification of *inherent uncertainty*, and (2) subsequently computes optimized perturbation parameters based on the determined *inherent uncertainty* levels to minimize perturbation errors while guaranteeing ε-LDP. Algorithm 1 presents the high-level framework.

4.4 Inherent Uncertainty Quantification

4.4.1 *Intuition.*

Applying a standard LDP mechanism to perturb the collected non-sensitive attribute A provides strong protection for the uncollected, correlated sensitive attribute S, but can be conservative for the utility of A. This is because it typically employs a fixed, worst-case perturbation strength that does not take into account the inherent uncertainty between A and S. We introduce parameters α and β to quantify this uncertainty; the ratio $\frac{\alpha}{\beta}$ captures the indirect privacy risk from A to S and allows us to tailor the perturbation strength: higher risk (larger $\frac{\alpha}{\beta}$) triggers stronger perturbation, while lower

Algorithm 1 Enhanced Pure LDP Protocols

Require: n: number of users; $\{v\}$: users' data; Ω_A, Ω_S: domains of correlated attribute A and sensitive attribute S; $\Pr(A|S)$: conditional distribution; ε: privacy budget

Ensure: the estimation of the true counts of $i \in \Omega_A$

1: the curator determines inherent uncertainty parameter (α, β) according to $\Pr(A|S)$;
2: the curator recalibrates perturbation parameter (p, q) according to (α, β) and ε;
3: perturb and aggregate the values of attribute A following the Base Protocols;
4: **return** the estimation of the true counts of $i \in \Omega_A$;

risk permits milder noise. As shown in the following sections, this risk-aware calibration preserves the same ε-LDP guarantee for S as the standard treat-as-sensitive baseline, while achieving higher utility for the collected data A.

4.4.2 *Uncertainty Parameter Selection Rule Definition.*

As we known, for any $\tilde{a}_k \in \Omega_{\tilde{A}}$ and $s_i \in \Omega_S$,

$$\Pr\left(\tilde{A} = \tilde{a}_k | S = s_i\right) = p^\circ \sum_{a \in supp(\tilde{a}_k)} \Pr(A = a | S = s_i) + q^\circ \left(1 - \sum_{a \in supp(\tilde{a}_k)} \Pr(A = a | S = s_i)\right), \quad (6)$$

where $supp(\tilde{a}_k)$ denotes the support set of $\tilde{a}_k$, and p°, q° are perturbation parameters which are defined in Equation 1. Let $cpd_{k|i} = \sum_{a \in supp(\tilde{a}_k)} \Pr(A = a | S = s_i)$. Furthermore, for any $\tilde{a}_k \in \Omega_{\tilde{A}}$ and $s_i, s_j \in \Omega_S$, we have

$$\frac{\Pr\left(\tilde{A} = \tilde{a}_k | S = s_i\right)}{\Pr\left(\tilde{A} = \tilde{a}_k | S = s_j\right)} = \frac{p^\circ \cdot cpd_{k|i} + q^\circ \cdot \left(1 - cpd_{k|i}\right)}{p^\circ \cdot cpd_{k|j} + q^\circ \cdot \left(1 - cpd_{k|j}\right)}.$$

Because p° and q° can be used to quantify *perturbation uncertainty*, we select an appropriate parameter pair from $\left\{\left(cpd_{k|i}, cpd_{k|j}\right) \mid \tilde{a}_k \in \Omega_{\tilde{A}}, s_i, s_j \in \Omega_S\right\}$ to quantify the *inherent uncertainty*. The selection rules can be defined as:

Definition 4.1 (Uncertainty Parameter Selection Rules). Given any two candidate parameter pairs $\left(cpd_{k_1|i_1}, cpd_{k_1|j_1}\right)$ and $\left(cpd_{k_2|i_2}, cpd_{k_2|j_2}\right)$ under ε-differential privacy, where $\tilde{a}_{k_1}, \tilde{a}_{k_2} \in \Omega_{\tilde{A}}$ and $s_{i_1}, s_{j_1}, s_{i_2}, s_{j_2} \in \Omega_S$, and with $cpd_{k_1|i_1} \geq cpd_{k_1|j_1}$ and $cpd_{k_2|i_2} \geq cpd_{k_2|j_2}$, the optimal parameter pair can be selected as follows. In particular, we allow $\tilde{a}_{k_1} = \tilde{a}_{k_2}$, $s_{i_1} = s_{i_2}$ or $s_{j_1} = s_{j_2}$.

- **Rule 1:** If $cpd_{k_1|i_1} \geq cpd_{k_2|i_2}$ and $cpd_{k_1|j_1} \leq cpd_{k_2|j_2}$, select $(cpd_{k_1|i_1}, cpd_{k_1|j_1})$.
- **Rule 2:** When $cpd_{k_1|i_1} > cpd_{k_2|i_2}$ and $cpd_{k_1|j_1} > cpd_{k_2|j_2}$, select $(cpd_{k_1|i_1}, cpd_{k_1|j_1})$ if $\frac{\Delta_{ki}}{\Delta_{kj}} \geq e^\varepsilon$, where $\Delta_{ki} = cpd_{k_1|i_1} - cpd_{k_2|i_2}$ and $\Delta_{kj} = cpd_{k_1|j_1} - cpd_{k_2|j_2}$.
- **Rule 3:** When $cpd_{k_1|i_1} > cpd_{k_2|i_2}$ and $cpd_{k_1|j_1} > cpd_{k_2|j_2}$, select $(cpd_{k_2|i_2}, cpd_{k_2|j_2})$ if $\frac{\Delta_{ki}}{\Delta_{kj}} < e^\varepsilon$.

We denote the ultimately selected parameter pair as (α, β), and employ it to quantify the *inherent uncertainty*.

4.4.3 *Correctness Analysis.*

The following theorem formalizes the accuracy of our *inherent uncertainty* parameter selection.

Theorem 4.2. *After selecting the optimal parameter pair (α, β), let*

$$\frac{p^\circ \cdot \alpha + q^\circ \cdot (1 - \alpha)}{p^\circ \cdot \beta + q^\circ \cdot (1 - \beta)} \leq e^\varepsilon.$$

Then, for any $\tilde{a}_k \in \Omega_{\tilde{A}}$ and $s_i, s_j \in \Omega_S$, we have

$$\frac{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_i\right)}{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_j\right)} = \frac{p^\circ \cdot cpd_{k|i} + q^\circ \cdot \left(1-cpd_{k|i}\right)}{p^\circ \cdot cpd_{k|j} + q^\circ \cdot \left(1-cpd_{k|j}\right)} \leq e^\varepsilon.$$

Proof. As (α, β) is the optimal parameter pair, for any another parameter pair $\left(cpd_{k|i}, cpd_{k|j}\right)$, exactly one of the following mutually exclusive conditions holds:

- **Case 1** ($\alpha \geq cpd_{k|i}$ and $\beta \leq cpd_{k|j}$): Rule 1 applies, and

$$\begin{aligned}\frac{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_i\right)}{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_j\right)} =& \frac{p^\circ \cdot cpd_{k|i} + q^\circ \cdot \left(1-cpd_{k|i}\right)}{p^\circ \cdot cpd_{k|j} + q^\circ \cdot \left(1-cpd_{k|j}\right)} = \frac{q^\circ + (p^\circ - q^\circ)\cdot cpd_{k|i}}{q^\circ + (p^\circ - q^\circ)\cdot cpd_{k|j}} \\ \leq& \frac{q^\circ + (p^\circ - q^\circ)\cdot \alpha}{q^\circ + (p^\circ - q^\circ)\cdot \beta} = \frac{p^\circ \cdot \alpha + q^\circ \cdot (1-\alpha)}{p^\circ \cdot \beta + q^\circ \cdot (1-\beta)} \leq e^\varepsilon\end{aligned}$$

- **Case 2** ($\alpha > cpd_{k|i}$, $\beta > cpd_{k|j}$, and $\frac{\Delta_\alpha}{\Delta_\beta} \geq e^\varepsilon$, where $\Delta_\alpha = \alpha - cpd_{k|i}, \Delta_\beta = \beta - cpd_{k|j}$): Rule 2 applies, and

$$\begin{aligned}\frac{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_i\right)}{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_j\right)} =& \frac{q^\circ + (p^\circ - q^\circ)\cdot cpd_{k|i}}{q^\circ + (p^\circ - q^\circ)\cdot cpd_{k|j}} = \frac{q^\circ + (p^\circ - q^\circ)\cdot \left(\alpha - \left(\alpha - cpd_{k|i}\right)\right)}{q^\circ + (p^\circ - q^\circ)\cdot \left(\beta - \left(\beta - cpd_{k|j}\right)\right)} \\ =& \frac{q^\circ + (p^\circ - q^\circ)\cdot (\alpha - \Delta_\alpha)}{q^\circ + (p^\circ - q^\circ)\cdot \left(\beta - \Delta_\beta\right)} = \frac{\left(q^\circ + (p^\circ - q^\circ)\cdot \alpha\right) - (p^\circ - q^\circ)\cdot \Delta_\alpha}{\left(q^\circ + (p^\circ - q^\circ)\cdot \beta\right) - (p^\circ - q^\circ)\cdot \Delta_\beta}\end{aligned}$$

  Because $\frac{q^\circ + (p^\circ - q^\circ)\cdot \alpha}{q^\circ + (p^\circ - q^\circ)\cdot \beta} \leq e^\varepsilon$ and $\frac{\Delta_\alpha}{\Delta_\beta} \geq e^\varepsilon$, we learn $\frac{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_i\right)}{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_j\right)} \leq e^\varepsilon$.
- **Case 3** ($\alpha < cpd_{k|i}$, $\beta < cpd_{k|j}$, and $\frac{\Delta_\alpha}{\Delta_\beta} < e^\varepsilon$, where $\Delta_\alpha = cpd_{k|i} - \alpha, \Delta_\beta = cpd_{k|j} - \beta$): Rule 3 applies, and following an analogous argument to Case 2, we derive $\frac{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_i\right)}{\Pr\left(\tilde{A}=\tilde{a}_k|S=s_j\right)} \leq e^\varepsilon$.

The proof is now complete. □

4.5 Adjusted Random Response Protocol (ARR)

4.5.1 *Inherent Uncertainty Parameter Selection.*

In adjusted random response protocol (ARR), both the input and output domain size is 2, i.e., $\Omega_A = \Omega_{\tilde{A}} = \{0, 1\}$, and $supp\left(\widetilde{a}=0\right) = \{0\}$ and $supp\left(\widetilde{a}=1\right) = \{1\}$. We determine the uncertain parameter pair (α, β) based on the conditional distribution $\Pr\left(A|S\right)$ as follows:

(1) For any $a_k \in \Omega_A$, we let

$$\alpha_k = \max\left\{\Pr(a_k|s_i)|s_i \in \Omega_S\right\}, \beta_k = \min\left\{\Pr(a_k|s_i)|s_i \in \Omega_S\right\}.$$

(2) For $a_1 = 0$ and $a_2 = 1$, if $\frac{\alpha_1}{\beta_1} \geq \frac{\alpha_2}{\beta_2}$, let $(\alpha, \beta) = (\alpha_1, \beta_1)$; otherwise, let $(\alpha, \beta) = (\alpha_2, \beta_2)$.

The correctness can be proved as follows:

(1) In Step 1, for any $s_i, s_j \in S$, the condition $\frac{\Pr(a_k|s_i)}{\Pr(a_k|s_j)} \leq \frac{\alpha_k}{\beta_k}$ holds, therefore the selection complies with Rule 1.
(2) In Step 2, for any $s \in \Omega_S$, we have $\Pr(A=0|S=s) + \Pr(A=1|S=s) = 1$. Thus, if

$$s_i = \arg\max_s \left\{\Pr\left(A=0|S=s\right)|s \in \Omega_S\right\}, \text{ and } s_j = \arg\min_s \left\{\Pr\left(A=0|S=s\right)|s \in \Omega_S\right\},$$

we have

$$s_i = \arg\min_s \{\Pr(A=1|S=s) \,|s \in \Omega_S\}, \text{ and } s_j = \arg\max_s \{\Pr(A=1|S=s) \,|s \in \Omega_S\}.$$

Thus, $\alpha_1 + \beta_2 = \beta_1 + \alpha_2 = 1$, and $\frac{\alpha_1 - \alpha_2}{\beta_1 - \beta_2} = 1 < e^\varepsilon$. Without loss of generality, if $\frac{\alpha_1}{\beta_1} \geq \frac{\alpha_2}{\beta_2}$, we can derive $\alpha_1 \leq \alpha_2, \beta_1 \leq \beta_2$. Therefore the selection complies with Rule 3.

4.5.2 *Perturbation Parameter Recalibration.*

Users perturb the values of attribute A with the perturbation function:

$$\Pr\left(\widetilde{A} = \widetilde{a} | A = a\right) = \begin{cases} p, & \text{if } \widetilde{a} = a \\ q, & \text{if } \widetilde{a} \neq a \end{cases},$$

Here, the parameters p and q satisfy:

$$p + q = 1, \tag{7}$$

$$\frac{\alpha \cdot p + (1-\alpha) \cdot q}{\beta \cdot p + (1-\beta) \cdot q} \leq e^\varepsilon. \tag{8}$$

Particularly, when $\frac{\alpha}{\beta} \leq e^\varepsilon$, we believe that no sensitive information about attribute S can be inferred from attribute A, so there is no need to perturb A. Therefore, we directly set $p = 1$ and $q = 0$. When $\frac{\alpha}{\beta} > e^\varepsilon$, we calculate p according to Equation 8 as follows:

$$p \leq \frac{(1-\alpha) - (1-\beta)\, e^\varepsilon}{(1-2\alpha) - (1-2\beta)\, e^\varepsilon}.$$

For a accurate analysis, we derive:

$$\frac{e^\varepsilon}{e^\varepsilon + 1} < p \leq \frac{(1-\alpha) - (1-\beta)\, e^\varepsilon}{(1-2\alpha) - (1-2\beta)\, e^\varepsilon} < 1.$$

4.5.3 *Utility Analysis.*

To evaluate the data utility of the ARR protocol, we compare it with that of the RR protocol. Recall that, in the traditional RR protocol, the perturbation parameters satisfy

$$\Pr\left(\widetilde{A} = \widetilde{a} | A = a\right) = \begin{cases} p' = \frac{e^\varepsilon}{e^\varepsilon + 1}, & \text{if } \widetilde{a} = a \\ q' = 1 - p' = \frac{1}{e^\varepsilon + 1}, & \text{if } \widetilde{a} \neq a \end{cases}.$$

Let var_{rr} and var_{arr} denote the variance of the estimated value in the RR protocol and ARR protocol, respectively. Since $p > \frac{e^\varepsilon}{e^\varepsilon + 1} = p'$ and $p + q = 1, p' + q' = 1$, it follows that $q < q'$. Thus, we have:

$$\text{var}_{\text{arr}} = \frac{nq(1-q)}{(p-q)^2} < \frac{nq'(1-q')}{(p'-q')^2} = \text{var}_{\text{rr}}.$$

Therefore, the data utility of the ARR protocol is superior to that of the RR protocol.

4.6 Adjusted GRR Protocol (AGRR)

4.6.1 *Inherent Uncertainty Parameter Selection.*

In adjusted GRR protocol (AGRR), we determine the uncertain parameter pair (α, β) as follows:

(1) For any $a_k \in \Omega_A$, we let

$$\alpha_k = \max\{\Pr(a_k|s_i)|s_i \in \Omega_S\}, \beta_k = \min\{\Pr(a_k|s_i)|s_i \in \Omega_S\}.$$

(2) Following the *Uncertainty Parameter Selection Rules*, we extract (α, β) from set $\{(\alpha_k, \beta_k) \mid a_k \in \Omega_A\}$

4.6.2 *Perturbation Parameter Recalibration.*

Users perturb the values of attribute A with the perturbation function:

$$\Pr(\widetilde{A} = \widetilde{a} | A = a) = \begin{cases} p, & \text{if } a = \widetilde{a} \\ q, & \text{if } a \neq \widetilde{a} \end{cases}.$$

Here, the parameters p and q satisfy the following conditions:

$$p + (d-1)q = 1, \text{ and } \frac{\alpha \cdot p + (1-\alpha) \cdot q}{\beta \cdot p + (1-\beta) \cdot q} \leq e^{\varepsilon}.$$

Similar to ARR, when $\frac{\alpha}{\beta} \leq e^{\varepsilon}$, we directly set $p = 1$ and $q = 0$. When $\frac{\alpha}{\beta} > e^{\varepsilon}$, we calculate p as follows:

$$p \leq \frac{(1-\alpha) - (1-\beta)e^{\varepsilon}}{(1-d\alpha) - (1-d\beta)e^{\varepsilon}}.$$

For a accurate analysis, we conclude that:

$$\frac{e^{\varepsilon}}{e^{\varepsilon} + d - 1} < p \leq \frac{(1-\alpha) - (1-\beta)e^{\varepsilon}}{(1-d\alpha) - (1-d\beta)e^{\varepsilon}} < 1.$$

4.6.3 *Utility Analysis.*

To evaluate the data utility of the AGRR protocol, we compare it with that of the GRR protocol. Similar to ARR, we can prove that the data utility of the AGRR protocol is superior to that of the GRR protocol.

4.7 Adjusted OUE Protocol (AOUE)

4.7.1 *Inherent Uncertainty Parameter Selection.*

Like OUE [32], in the AOUE protocol, a perturbed output may support multiple inputs. Therefore, when determining the uncertainty parameters, we need to consider the corresponding subset of the domain of attribute A. The specifics are outlined below:

(1) For any given $s_i, s_j \in \Omega_S$, we construct the set $\mathbb{C}_{ij}$,

$$\mathbb{C}_{ij} = \left\{ a \mid a \in \Omega_A \wedge \frac{\Pr\left(A = a \mid S = s_i\right)}{\Pr\left(A = a \mid S = s_j\right)} \geq e^{\varepsilon} \right\}.$$

Then, we calculate α_{ij} and β_{ij} as:

$$\alpha_{ij} = \sum_{a \in \mathbb{C}_{ij}} \Pr\left(A = a \mid S = s_i\right), \beta_{ij} = \sum_{a \in \mathbb{C}_{ij}} \Pr\left(A = a \mid S = s_j\right).$$

Note that, if $\mathbb{C}_{ij} = \emptyset$, let $\alpha_{ij} = \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$, $\beta_{ij} = \frac{1}{1+e^{\varepsilon}}$.

(2) Following the *Uncertainty Parameter Selection Rules*, we extract (α, β) from set $\left\{(\alpha_{ij}, \beta_{ij}) \mid s_i, s_j \in S\right\}$.

To verify the correctness of parameter (α, β) selection, it suffices to demonstrate that the choice of $(\alpha_{ij}, \beta_{ij})$ for each (s_i, s_j) complies with the selection rules. Specifically, for each perturbed vector $\tilde{B}$, denote

$$U^+ = C_{ij} - supp\left(\tilde{B}\right), \quad U^- = supp\left(\tilde{B}\right) - C_{ij}, \text{ and}$$

$$\Delta_{\alpha 1} = \sum_{a \in U^+} \Pr\left(A = a | S = s_i\right), \Delta_{\alpha 2} = \sum_{a \in U^-} \Pr\left(A = a | S = s_i\right)$$

$$\Delta_{\beta 1} = \sum_{a \in U^+} \Pr\left(A = a | S = s_j\right), \Delta_{\beta 2} = \sum_{a \in U^-} \Pr\left(A = a | S = s_j\right)$$

Then, we can learn that,

$$\sum_{a\in supp(\tilde{B})} \Pr\left(A=a|S=s_i\right)=\alpha_{ij}-\Delta_{\alpha 1}+\Delta_{\alpha 2}, \text{ and } \sum_{a\in supp(\tilde{B})} \Pr\left(A=a|S=s_j\right)=\beta_{ij}-\Delta_{\beta 1}+\Delta_{\beta 2}$$

when $U^+\neq\emptyset$ and $U^-\neq\emptyset$, $\frac{\Delta_{\alpha 1}}{\Delta_{\beta 1}}\geq e^\varepsilon$ and $\frac{\Delta_{\alpha 2}}{\Delta_{\beta 2}}<e^\varepsilon$. Thus the selection of (α_{ij},β_{ij}) complies with Rules 2 and 3.

4.7.2 Perturbation Parameter Recalibration.

In AOUE, users perform bit-wise perturbation on the encoded vectors. The perturbation parameters satisfy:

$$\Pr\left(\widetilde{B}[l]=1\right)=\begin{cases}p, & \text{if } B[l]=1\\ q, & \text{if } B[l]=0\end{cases}.$$

where the parameters p and q satisfy

$$\frac{\alpha p(1-q)+(1-\alpha)(1-p)q}{\beta p(1-q)+(1-\beta)(1-p)q}\leq e^\varepsilon. \tag{9}$$

Specifically, when $\frac{\alpha}{\beta}\leq e^\varepsilon$, we directly set $p=1$ and $q=0$. When $\frac{\alpha}{\beta}>e^\varepsilon$, we calculate the optimal perturbation parameters p and q according to Lemma 4.3.

Lemma 4.3. *When $\frac{\alpha}{\beta}>e^\varepsilon$, the perturbation parameters of AOUE can be set to be:*

$$p=\frac{1}{2},\ \textit{and}\ q=\frac{\alpha-\beta e^\varepsilon}{(e^\varepsilon-1)+2(\alpha-\beta e^\varepsilon)}.$$

Proof. The proof is omitted due to space limitations. □

4.7.3 Utility Analysis.

When $\frac{\alpha}{\beta}\leq e^\varepsilon$, $p=1, q=0$, $var=\frac{nq(1-q)}{(p-q)^2}=0$. When $\frac{\alpha}{\beta}>e^\varepsilon$, according to Lemma 4.3, the parameters p and q of AOUE are set to be $p=\frac{1}{2}$, and $q=\frac{\alpha-\beta e^\varepsilon}{(e^\varepsilon-1)+2(\alpha-\beta e^\varepsilon)}$. In the OUE protocol, $p_{oue}=\frac{1}{2}$, $q_{oue}=\frac{1}{e^\varepsilon+1}<\frac{1}{2}$. In addition, it is known that $\alpha-\beta e^\varepsilon<1$. So,

$$q=\frac{\alpha-\beta e^\varepsilon}{(e^\varepsilon-1)+2(\alpha-\beta e^\varepsilon)}<\frac{1}{e^\varepsilon+1}=q_{oue}.$$

Therefore,

$$\text{var}=\frac{nq(1-q)}{(p-q)^2}<\frac{nq_{oue}(1-q_{oue})}{(p_{oue}-q_{oue})^2}=\text{var}_{oue}.$$

5 Hybrid Privacy Risk Prevention

In this section, we present a systematic approach to the hybrid privacy risk problem through representative missing data collection. We begin by formally defining the problem, followed by proposing a two-stage perturbation framework accompanied by rigorous privacy-utility trade-off analysis.

5.1 Motivating Examples

Consider a scenario where a data collector aims to gather users' *cholesterol level* (Attribute A). This collection process encounters a *hybrid privacy risk* problem: (1) Direct Risk: A cholesterol reading like "250 mg/dL" is sensitive. Disclosing it directly violates privacy. (2) Indirect Risk: For some users, this data is missing. The missing status (e.g., recording "NULL") is itself non-sensitive information that must be collected. However, if this missingness is correlated with a sensitive attribute like *age* (e.g., missing rates are high for the very young and the elderly), an adversary can exploit this link

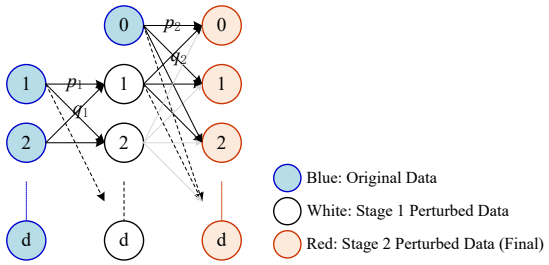


Fig. 1. **Workflow of the proposed two-stage perturbation method. Visual Element: Blue circles:** represent original user data (including valid and missing values); **White circles** denote intermediate, valid results after Stage 1 perturbation; **Red circles** indicate the final perturbed results after Stage 2 perturbation, which applies to all values.

to infer a user's age group from the missing data pattern that is reported without perturbation. This example illustrates the core challenge we address: protecting against the direct risk of sensitive values of A and the indirect risk of an uncollected sensitive attribute S *simultaneously* during the collection of A.

5.2 Problem Definition

Problem 2. *There are n users, each possessing one data point containing some attributes. Among them, two sensitive attributes are denoted as S and A, with domains Ω_S and Ω_A, respectively. For some users among the n users, the value of attribute A is missing, and its missing status is related to S. Let M represent the missing status of attribute A, with the domain $\Omega_M = \{m_1, m_2\}$, where m_1 indicates missing and m_2 indicates non-missing. The conditional distribution $\Pr(M|S)$ is known a priori. Now, an aggregator needs to collect frequency information of attribute A in users' data while satisfying ε-LDP:*

(1) ***Indirect privacy risk prevention****. For any perturbed missing status $\widetilde{m} \in \Omega_{\widetilde{M}}$ and $s_i, s_j \in \Omega_S$:*

$$\Pr\left(\widetilde{M} = \widetilde{m} | S = s_i\right) \leq e^{\varepsilon} \cdot \Pr\left(\widetilde{M} = \widetilde{m} | S = s_j\right). \tag{10}$$

(2) ***Direct privacy risk prevention****. For any perturbed valid value $\widetilde{a} \in \Omega_{\widetilde{A}}$ and $a_i, a_j \in \Omega_A$:*

$$\Pr\left(\widetilde{A} = \widetilde{a} | A = a_i\right) \leq e^{\varepsilon} \cdot \Pr\left(\widetilde{A} = \widetilde{a} | A = a_j\right). \tag{11}$$

5.3 Two-Stage Perturbation Method

5.3.1 Protocol Process.

Perturbation. Fig. 1 provides a visual overview of the workflow for our proposed two-stage perturbation method, illustrating how data evolves from its original state, through an intermediate perturbation, to the final output. In the method, we first introduce an "equivalent" amount of uncertainty to the valid values according to the *inherent uncertainty* of the missing status, and then apply a unified perturbation to both the missing value and the perturbed valid values from the first stage. Alg. 2 describes the details of the Two-Stage Perturbation Method for missing data collection.

Specifically, in Stage 1, we perturb the valid values of A (denoted as $1, \ldots, d$) with parameters:

$$\Pr(\mathcal{P}(i) = j) = \begin{cases} p_1, & \text{if } i = j, \\ q_1, & \text{if } i \neq j. \end{cases}$$

Algorithm 2 Two-Stage Perturbation Method

Require: n: number of users; $\{v\}$: users' data; $\Omega_A = \{1, \ldots, d\}$, Ω_S: domains of A and S; Ω_M: domain of missing status of A; $\Pr(M|S)$: conditional distribution; ε: privacy budget

Ensure: estimation of counts of valid and missing values of A

▷ Inherent uncertainty parameter selection by the collector

learn (α, β) from $\Pr(M|S)$ according to Selection Rules;

▷ Perturbation parameter calculation by the collector

$p_2 = \frac{(1-\alpha)-(1-\beta)e^{\varepsilon}}{(1-(d+1)\alpha)-(1-(d+1)\beta)e^{\varepsilon}}$, $q_2 = \frac{1-p_2}{d}$;

$p_1 = \frac{e^{\varepsilon}p_2+((d-2)e^{\varepsilon}-(d-1))q_2}{(d-1+e^{\varepsilon})(p_2-q_2)}$, $q_1 = \frac{1-p_1}{d-1}$;

▷ Stage 1 perturbation, $i, j \in \Omega_A = \{1, \ldots, d\}$

for each user with valid value **do**

perturb the valid value of A with the perturbation $\mathcal{P}_1$:

$$\Pr(\mathcal{P}_1(i) = j) = \begin{cases} p_1, & \text{if } i = j \\ q_1, & \text{if } i \neq j \end{cases};$$

end for

▷ Stage 2 perturbation, $i, j \in \{0, 1, \ldots, d\}$, 0 is missing value

for each user **do**

perturb the data with the perturbation $\mathcal{P}_2$:

$$\Pr(\mathcal{P}_2(i) = j) = \begin{cases} p_2, & \text{if } i = j \\ q_2, & \text{if } i \neq j \end{cases};$$

end for

▷ Estimation of valid and missing value counts by the collector

obtain the perturbed count c_i of $i \in \{0, \ldots, d\}$;

$p^{\star} = p_1 \cdot p_2 + (1 - p_1) \cdot q_2$, $q^{\star} = q_1 \cdot p_2 + (1 - q_1) \cdot q_2$;

$\widetilde{n}_0 = \frac{c_0 - n \cdot q_2}{p_2 - q_2}$, $\widetilde{n}_i = \frac{c_i - \widetilde{n}_0 \cdot q_2 - (n - \widetilde{n}_0) \cdot q^{\star}}{p^{\star} - q^{\star}}$, where $1 \leq i \leq d$;

return $\widetilde{n}_0, \widetilde{n}_1, \ldots, \widetilde{n}_d$;

where $i, j \in \{1, \ldots, d\}$. In Stage 2, we simultaneously perturb both the missing value (treated as 0) and the perturbed valid values from the first stage (i.e., 1,…,d) with parameters:

$$\Pr(\mathcal{P}(i) = j) = \begin{cases} p_2, & \text{if } i = j, \\ q_2, & \text{if } i \neq j. \end{cases}$$

where $i, j \in \{0, 1, \ldots, d\}$.

In particular, the parameters satisfy:

$$p_1 + (d-1) \cdot q_1 = 1, \qquad p_2 + d \cdot q_2 = 1,$$
$$\frac{\alpha \cdot p_2 + (1-\alpha) \cdot q_2}{\beta \cdot p_2 + (1-\beta) \cdot q_2} \leq e^{\varepsilon}, \quad \frac{p_1 \cdot p_2 + (1-p_1) \cdot q_2}{q_1 \cdot p_2 + (1-q_1) \cdot q_2} \leq e^{\varepsilon}.$$

Because the domain size of missing status is 2, i.e., $|\Omega_M| = 2$, the inherent uncertainty parameter pair (α, β) can be selected as described in Section 4.5.1, i.e.,

$$\alpha = \max\{\Pr(M = m^*|S = s) \mid s \in \Omega_S\}, \text{ and } \beta = \min\{\Pr(M = m^*|S = s) \mid s \in \Omega_S\},$$

where $m^* = \arg\max_{m \in \Omega_M} \left\{ \frac{\max\{\Pr(M=m|S=s)|s \in \Omega_S\}}{\min\{\Pr(M=m|S=s)|s \in \Omega_S\}} \right\}$.

When $\frac{\alpha}{\beta} \leq e^{\varepsilon}$, we directly set $p_2 = 1, q_2 = 0$. When $\frac{\alpha}{\beta} > e^{\varepsilon}$, we calculate p_2 as:

$$p_2 \leq \frac{(1-\alpha) - (1-\beta)e^{\varepsilon}}{(1-(d+1)\alpha) - (1-(d+1)\beta)e^{\varepsilon}}.$$

For a accurate analysis, we derive:

$$\frac{e^{\varepsilon}}{e^{\varepsilon}+d} < p_2 \leq \frac{(1-\alpha) - (1-\beta)e^{\varepsilon}}{(1-(d+1)\alpha) - (1-(d+1)\beta)e^{\varepsilon}} < 1.$$

Aggregation. The aggregator collects the users' perturbed data and obtains the counts of missing value and valid values after perturbation, denoted as $c_0, \ldots, c_d$.
Estimation. Denote the original input value as in and the output value as out. When $out = 0$, we can get:

$$\Pr(out = 0 | in = 0) = p_2, \text{ and } \Pr(out = 0 | in = i \neq 0) = q_2.$$

When $out \neq 0$,

$$\Pr(out = j | in = 0) = q_2,$$
$$\Pr(out = j | in = j) = p_1 \cdot p_2 + (1-p_1) \cdot q_2.$$

When the input value is different from the output value and the input value is not 0, i.e., $i \neq j \wedge i \neq 0$,

$$\Pr(out = j | in = i) = q_1 \cdot p_2 + (1-q_1) \cdot q_2.$$

Let $p^{\star} = p_1 \cdot p_2 + (1-p_1) \cdot q_2$ and $q^{\star} = q_1 \cdot p_2 + (1-q_1) \cdot q_2$, and the true counts be $n_0, \ldots, n_d$. According to maximum likelihood estimation, we obtain,

$$n_0 \cdot p_2 + (n - n_0) \cdot q_2 \approx c_0.$$

When $1 \leq i \leq d$,

$$n_0 \cdot q_2 + n_i \cdot p^{\star} + (n - n_0 - n_i) \cdot q^{\star} \approx c_i.$$

Therefore, we obtain the estimated counts as:

$$\widetilde{n}_0 = \frac{c_0 - n \cdot q_2}{p_2 - q_2}, \text{ and } \widetilde{n}_i = \frac{c_i - \widetilde{n}_0 \cdot q_2 - (n - \widetilde{n}_0) \cdot q^{\star}}{p^{\star} - q^{\star}}.$$

5.3.2 *Privacy Analysis.*

Theorem 5.1. *The two-stage perturbation scheme satisfies ε-LDP.*

Proof. To prove that the two-stage scheme satisfies ε-LDP, we need to prove that both Eq. 10 and Eq. 11 hold. We begin by proving the validity of Eq. 10 for cases where the output is a missing value. Let m_1 and m_2 represent the missing and non-missing states, respectively. Without loss of generality, let

$$m_1 = \arg\max_{m \in \Omega_M} \left\{ \frac{\max_{s \in \Omega_S} \Pr(M = m | S = s)}{\max_{s \in \Omega_S} \Pr(M = m | S = s)} \right\}.$$

Then, for any $s_i \in S$, $\alpha \leq \Pr(M = m_1 | S = s_i) \leq \beta$. We can calculate the conditional probability

$$\Pr(\widetilde{M} = m_1 | S = s_i) = \Pr(M = m_1 | S = s_i) \cdot p_2 + (1 - \Pr(M = m_1 | S = s_i)) \cdot q_2.$$

Since $p_2 > q_2$ and $\beta \leq \Pr(M = m_1 | S = s_i) \leq \alpha$, we have

$$\beta \cdot p_2 + (1-\beta) \cdot q_2 \leq \Pr(\widetilde{M} = m_1 | S = s_i) \leq \alpha \cdot p_2 + (1-\alpha) \cdot q_2.$$

Therefore, for any $s_i, s_j \in S$,

$$\frac{\Pr(\widetilde{M} = m_1 | S = s_i)}{\Pr(\widetilde{M} = m_1 | S = s_j)} \leq \frac{\alpha \cdot p_2 + (1-\alpha) \cdot q_2}{\beta \cdot p_2 + (1-\beta) \cdot q_2} \leq e^{\varepsilon}.$$

Furthermore, we can prove that Eq. 10 still holds when the output value is a non-missing value m_2. As $\Pr(\widetilde{M}=m_2|S=s_i)=\sum_{k=1}^{d}\Pr(\widetilde{A}=k|S=s_i)$, we first calculate $\Pr(\widetilde{A}=k|S=s_i)$ for each $k\in[1,\dots,d]$. Specifically,

$$\Pr(\widetilde{A}=k|S=s_i)=\Pr(M=m_1|S=s_i)\cdot q_2+\Pr(M=m_2|S=s_i)\cdot\Pr(\widetilde{A}=k|M=m_2).$$

In the above equation,

$$\Pr(\widetilde{A}=k|M=m_2)=\frac{\Pr(\widetilde{A}=k,M=m_2)}{\Pr(M=m_2)}=\frac{\sum_{l=1}^{d}\Pr(A=l)\Pr(\widetilde{A}=k|A=l)}{\sum_{l'=1}^{d}\Pr(A=l')}$$
$$\geq\min\{\Pr(\widetilde{A}=k|A=l)\mid l\in\{1,\dots,d\}\}=\Pr(\widetilde{A}=k|A=l\neq k)=q^{\star}>q_2.$$

Thus, we have

$$(1-\alpha)\cdot\Pr(\widetilde{A}=k|M=m_2)+\alpha\cdot q_2\leq\Pr(\widetilde{A}=k|S=s_i)\leq(1-\beta)\cdot\Pr(\widetilde{A}=k|M=m_2)+\beta\cdot q_2.$$

Furthermore,

$$\Pr(\widetilde{A}=k|M=m_2)=\frac{\sum_{l=1}^{d}\Pr(A=l)\Pr(\widetilde{A}=k|A=l)}{\sum_{l'=1}^{d}\Pr(A=l')}$$
$$\leq\max\{\Pr(\widetilde{A}=k|A=l)\mid l\in\{1,\dots,d\}\}=\Pr(\widetilde{A}=k|A=k)=p^{\star}\leq p_2.$$

Let $\Delta=p_2-\Pr(\widetilde{A}=k|M=m_2)$. Therefore,

$$(1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2\leq\Pr(\widetilde{A}=k|S=s_i)\leq(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2.$$

Recall that $\frac{1-\beta}{1-\alpha}\leq\frac{\alpha}{\beta}$. When $\frac{1-\beta}{1-\alpha}\leq e^{\varepsilon}$, obviously,

$$\frac{\Pr(\widetilde{A}=k|S=s_i)}{\Pr(\widetilde{A}=k|S=s_j)}\leq\frac{(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2}{(1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2}\leq e^{\varepsilon}.$$

When $\frac{1-\beta}{1-\alpha}>e^{\varepsilon}$,

$$\frac{(1-\beta)\cdot p_2+\beta\cdot q_2}{(1-\alpha)\cdot p_2+\alpha\cdot q_2}-\frac{\alpha\cdot p_2+(1-\alpha)\cdot q_2}{\beta\cdot p_2+(1-\beta)\cdot q_2}=\frac{(\beta(1-\beta)-\alpha(1-\alpha))\cdot(p_2^2+q_2^2-2p_2q_2)}{((1-\alpha)\cdot p_2+\alpha\cdot q_2)\cdot(\beta\cdot p_2+(1-\beta)\cdot q_2)}\leq 0.$$

Therefore,

$$\frac{(1-\beta)\cdot p_2+\beta\cdot q_2}{(1-\alpha)\cdot p_2+\alpha\cdot q_2}\leq\frac{\alpha\cdot p_2+(1-\alpha)\cdot q_2}{\beta\cdot p_2+(1-\beta)\cdot q_2}\leq e^{\varepsilon}.$$

Furthermore, since

$$(1-\beta)\cdot p_2+\beta\cdot q_2\leq e^{\varepsilon}((1-\alpha)\cdot p_2+\alpha\cdot q_2),\ \text{and}\ -\Delta(1-\beta)<e^{\varepsilon}(-\Delta(1-\alpha)),$$

we have

$$(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2\leq e^{\varepsilon}((1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2).$$

Then, for any $s_i,s_j\in S$,

$$\frac{\Pr(\widetilde{A}=k|S=s_i)}{\Pr(\widetilde{A}=k|S=s_j)}\leq\frac{(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2}{(1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2}\leq e^{\varepsilon}.$$

Therefore,

$$\frac{\Pr(\widetilde{M}=m_2|S=s_i)}{\Pr(\widetilde{M}=m_2|S=s_j)}=\frac{\sum_{k=1}^{d}\Pr(\widetilde{A}=k|S=s_i)}{\sum_{k=1}^{d}\Pr(\widetilde{A}=k|S=s_j)}\leq e^{\varepsilon}.$$

Up to this point, we have proved that Eq. 10 holds for any $m_k \in \widetilde{M}, s_i, s_j \in S$. Then, we can prove that Eq. 11 holds by $\frac{\Pr(\widetilde{A}=\widetilde{a}|A=a_i)}{\Pr(\widetilde{A}=\widetilde{a}|A=a_j)} \leq \frac{p^\star}{q^\star} \leq e^\varepsilon$. Therefore, the two-stage scheme satisfies ε-LDP. □

5.3.3 Utility Analysis.

To evaluate the data utility of the two-stage scheme, we compare it with that of a unified perturbation scheme. The so-called unified perturbation ignores the *inherent uncertainty* of the missing value and directly perturbs the missing status (0) and valid values (1, …, d) with the probability:

$$\Pr(\mathcal{P}(i) = j) = \begin{cases} p = \frac{e^\varepsilon}{e^\varepsilon + d}, & \text{if } i = j \\ q = 1 - dp = \frac{1}{e^\varepsilon + d}, & \text{if } i \neq j \end{cases}.$$

where $i \in \{0, 1, \ldots, d\}$. In the unified perturbation scheme, the variance of the estimated value for both missing and valid values is $\text{var} = \frac{nq(1-q)}{(p-q)^2}$. In the two-stage scheme, the variance of the estimated missing value is $\text{var}_0 = \frac{nq_2(1-q_2)}{(p_2-q_2)^2}$. Since $p_2 > \frac{e^\varepsilon}{e^\varepsilon+d} = p$, and $p_2 + dq_2 = 1$, $p + dq = 1$, we obtain $q_2 < q$. Therefore,

$$\text{var}_0 = \frac{nq_2(1-q_2)}{(p_2-q_2)^2} < \frac{nq(1-q)}{(p-q)^2} = \text{var}.$$

In the two-stage perturbation scheme, the variance of the estimated value for any valid value is $\text{var}_i = \frac{nq^\star(1-q^\star)}{(p^\star-q^\star)^2}$. Since $p^\star + (d-1)q^\star + q_2 = 1$, $p + dq = 1$, and $\frac{p^\star}{q^\star} = \frac{p}{q} = e^\varepsilon$, then $q^\star > q$. Simultaneously,

$$\text{var}_i = \frac{nq^\star(1-q^\star)}{(p^\star-q^\star)^2} = \frac{n}{(e^\varepsilon-1)^2} \cdot \left(\frac{1}{q^\star} - 1\right), \text{ and } \text{var} = \frac{nq(1-q)}{(p-q)^2} = \frac{n}{(e^\varepsilon-1)^2} \cdot \left(\frac{1}{q} - 1\right).$$

Therefore, $\text{var}_i < \text{var}$, and the data utility of the two-stage perturbation scheme is superior to that of the unified perturbation scheme.

6 System Integration

6.1 System Architecture

Our method naturally fits a standard *Compute-Broadcast-Perturb* architecture for practical deployment. The data collector first performs a one-time inherent uncertainty quantification and perturbation parameter recalibration, and publishes perturbation parameters to a configuration service. Millions of user devices then fetch these parameters and perturb their local data in $O(1)$ time before transmission. This design centralizes the computational overhead into a single server-side pre-processing step, enabling nearly limitless client-side scalability.

From an implementation standpoint, the parameter computation can be packaged as a User-Defined Aggregate Function (UDAF) in systems, while the local perturbation is a lightweight library function in a client Software Development Kit (SDK). This architecture is particularly compelling for privacy-sensitive analytics at scale, such as collecting high-fidelity user engagement metrics from mobile apps for recommendation models while rigorously protecting correlated sensitive demographics.

The scheme offers an excellent utility-deployment trade-off: its asymmetric efficiency (one-time server cost vs. minimal client overhead) is its key advantage, with the primary practical consideration being the need for periodic parameter recomputation should the underlying conditional distribution $\Pr(A|S)$ shift significantly.

6.2 Efficiency and Scalability Analysis

Our approach introduces two key steps not found in standard LDP protocols: inherent uncertainty quantification and perturbation parameter recalibration. To address potential concerns about the overhead of these steps, our Compute-Broadcast-Perturb model is designed to execute them as a one-time computation solely on the server side. This architectural decision leads to the following formal overhead decomposition:

- Server-Side Overhead: The data collector performs a one-time computation to derive the perturbation parameters (p, q). The complexity of this step is independent of the number of users n, scaling only polynomially with the attribute domain sizes $|\Omega_A|$ and $|\Omega_S|$.
- Client-Side Overhead: Each user perturbs their data locally in $O(1)$ time, imposing a negligible burden on end-user devices.
- Communication Overhead: The system ensures optimal communication efficiency in both directions: 1) Downlink (Server to Users). The data collector broadcasts the final perturbation parameters, merely one or two pairs of floats, via an efficient one-to-many channel. This incurs a minimal per-user communication cost of $O(1)$. 2) Uplink (Users to Server). Crucially, even in the missing data scenario which requires two local perturbations, each user submits their data to the server in a single communication round. This maintains the $O(1)$ per-user uplink cost, matching the efficiency of standard LDP protocols and avoiding any additional communication cost.
- Storage Overhead: The server-side storage overhead is minimal and scales only with the data domain. Specifically, the server must store the inherent uncertainty information (i.e., the conditional probability distributions), whose size is a constant dependent on the attribute domain sizes and is independent of the number of users. This constitutes no practical storage burden compared to standard LDP protocols and is a key enabler for scalability.

This architecture centralizes the one-time, data-dependent cost on the server, ensuring efficient and scalable deployment for massive user bases.

7 Experiments

In this section, we provide a comprehensive experimental study of our approaches. Note that, to better evaluate the performance of our approaches, we have not only conducted tests on the real-world datasets, but also demonstrated the effectiveness of the approaches within a broader range of parameter values. All experiments are performed ten times and we plot the mean.

7.1 Experimental Setup

Methods: We evaluate our solutions by comparing them with LDP protocols that do not consider the *inherent uncertainty* of the data.

Correlation attribute collection.

- **Adjusted Protocols.** For the collection of correlation attribute, we propose optimization LDP protocols that fully leverage the *inherent uncertainty* of the data. These optimized protocols are named Adjusted Randomized Response (*ARR*), Adjusted Generalized Randomized Response (*AGRR*), and Adjusted Optimized Unary Encoding (*AOUE*).
- **Original protocols.** We compare their performance with that of the original protocols, namely *RR*, *GRR*, and *OUE*, respectively.

Perturbation Target and Comparison Fairness. For both our proposed adjusted protocols and all baseline methods (*RR*, *GRR*, and *OUE*), the perturbation is applied exclusively to the collected non-sensitive attribute A. The sensitive attribute S is neither collected nor perturbed in our problem setting. This is a fundamental distinction of our work compared to traditional LDP literature.

Table 1. The details of selected attributes.

methods	datasets	correlation/missing attribute		sensitive attributes	
		name	domain size	name	domain size
ARR	Adult	salary	2	sex	2
	Diabetes	NoDocbcCost	2	AnyHealthcare	2
AGRR/AOUE	Adult	age	100	salary	2
	Diabetes	BMI	90	HighChol	2
TSP	Adult	salary	2	sex	2
	Diabetes	GenHlth	5	AnyHealthcare	2

Our comparison aims to evaluate which perturbation mechanism, when applied to A, can more effectively defend against the inference of the correlated (and uncollected) sensitive attribute S, under the same privacy budget with respect to S.

Missing data collection.

- **TSP**. For the collection of missing data, we propose a two-stage perturbation scheme, which is referred to as TSP. TSP first applies a certain degree of perturbation to the valid values, and then, applies the same scale of perturbation to the valid values and the missing status.
- **UP**. In contrast, the baseline approach treats the missing status as a distinct attribute value and perturbs them together with other attribute values in a unified manner. We refer to this approach as Unified Perturbation (UP).
- **BiSample**. The existing work [29] proposes *BiSample*, a bi-directional sampling mechanism for data perturbation, to address the issue of missing data collection for mean estimation. We also compare with this method in our paper.

Datasets: We use the following two real datasets to conduct our experiments.

- *Adult*[1], is a dataset containing 45,222 records extracted by Barry Becker from the 1994 Census database. This dataset contains 13 attributes.
- *Diabetes*[2], is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. This dataset contains 22 attributes.

For each dataset, we select some attributes to conduct our experiments. For example, we select *salary* as correlation attribute, and *sex* as sensitive attribute from *Adult* to evaluate the performance of *ARR*. We collect information on attribute *salary* while preventing attackers from inferring the privacy of *sex* based on information about *salary*. The correlation between *salary* and *sex* is treated as *inherent uncertainty*. The details of selected attributes are described in Table 1.

Tasks and metrics: We employ the Mean Squared Error (MSE) of frequency estimaiton (referred to as f) as a metric to evaluate the performance of our methods, where MSE of estimated frequency can be calculated as:

$$MSE_f = \frac{1}{|\Omega_A|} \sum_{i \in \Omega_A} \left(f_i - \tilde{f}_i \right)^2 . \tag{12}$$

where Ω_A and $|\Omega_A|$ denote the attribute domain and domain size respectively, and $\left(f_i - \tilde{f}_i \right)^2$ denotes the squared error of the estimated frequency of the value i.

Note that, *TSP* and *UP* can support fine-grained frequency estimation, certainly including null value rate (referred to as *nr*), while *Bisample* only supports estimation of the mean (referred to as

[1]https://archive.ics.uci.edu/dataset/2/adult
[2]https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

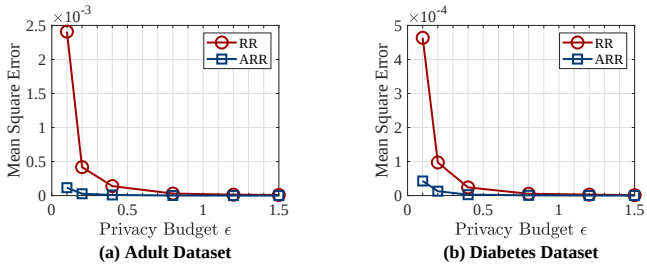


Fig. 2. MSE of *ARR*/*RR* over real datasets.

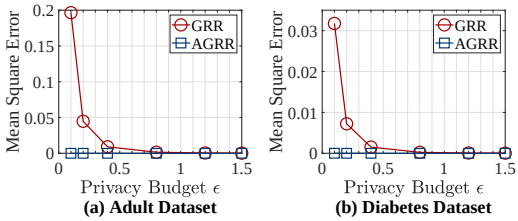


Fig. 3. MSE of AGRR/GRR over real datasets.

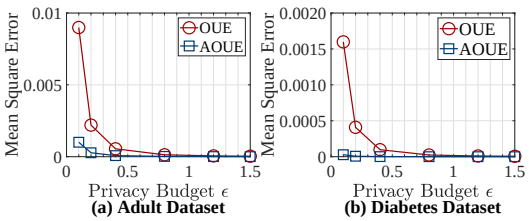


Fig. 4. MSE of AOUE/OUE over real datasets.

m) and null value rate. To compare with *Bisample*, we first estimate frequencies of valid values by using *TSP* and then calculate their mean. The Squared Error of mean can be calculated as:

$$SE_{\mathrm{m}} = \left(\sum_{i \in \Omega_A} i \cdot f_i - \sum_{i \in \Omega_A} i \cdot \tilde{f}_i \right)^2 \tag{13}$$

Parameters Setup. In the experiments, the involved parameters include the conditional probabilities α and β that determine the inherent uncertainty of the data, the overall privacy budget ε, and the size of the attribute domain d. We set $\beta \leq \alpha$. Specifically, for the *RR* and *ARR* protocols, $d = 2$. For the collection of missing data, we also need to consider the missing rate (referred to as δ).

7.2 Results

7.2.1 *Adjusted Protocols vs Original Protocols.*

We benchmark our protocols against the original LDP baselines (RR, GRR, OUE) on collecting the correlated non-sensitive attribute A, with all methods applied exclusively to A. The evaluation on two real datasets is followed by an analysis of their performance across a broad parameter range.

Figs. 2, 3, and 4 demonstrate MSE in frequency estimation for correlation attributes in both the adjusted protocols (ARR, AGRR, AOUE) and original protocols (RR, GRR, OUE) across different privacy budgets ε, under two real-world datasets. It can be observed that as ε increases, the MSE decreases for both kinds of protocols. This is due to the reduced overall privacy requirement, which leads to a decrease in the scale of perturbations for original protocols. Meanwhile, for the adjusted protocols, the uncertainty required by LDP also decreases. Furthermore, our proposed protocols consistently outperform the original protocols. The advantage is more pronounced when ε is lower. Additionally, we find that in real data, by accounting for inherent uncertainty, only a low level of noise needs to be introduced to satisfy ε-LDP.

Next, we demonstrate the effectiveness of the protocols within a broader range of parameter values. Figs. 5(a), 6(a), and 7(a) show the performance of the protocols under different conditional

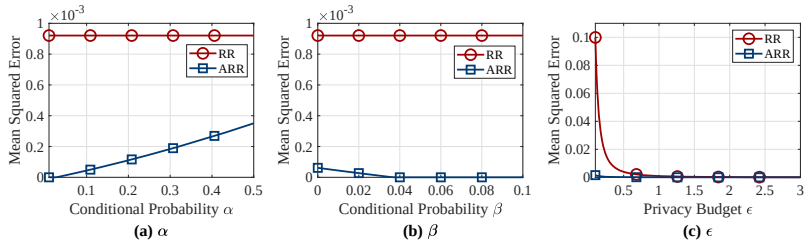


Fig. 5. MSE of *ARR*/*RR* within a broader range of parameter value.

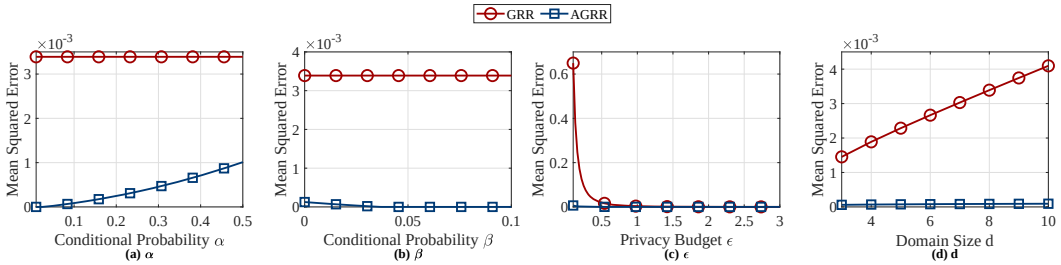


Fig. 6. MSE of *AGRR*/*GRR* within a broader range of parameter value.

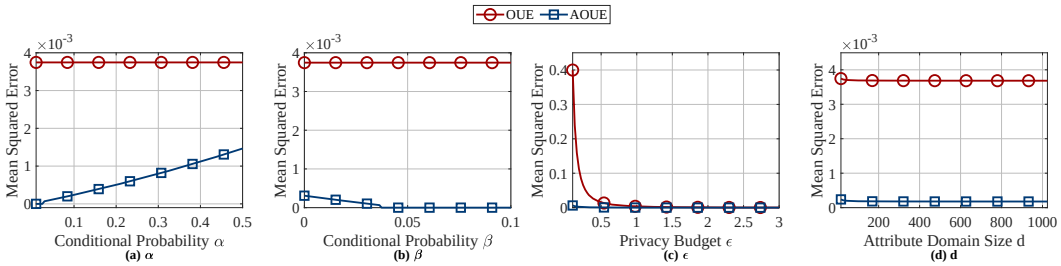


Fig. 7. MSE of *AOUE*/*OUE* within a broader range of parameter value.

probability α. We set $\varepsilon = 1$ and $\beta = 0.01$. The range of α is defined as $[\beta, 0.5)$. It can be observed that, adjusted protocols consistently achieve a lower MSE compared to original protocols, and as α increases, the MSE of adjusted protocols continuously increases. This is due to a decrease in the *inherent uncertainty*, resulting in an increase in *perturbation uncertainty*.

Figs. 5(b), 6(b), and 7(b) show the performance of the protocols under different conditional probability β. We set $\varepsilon = 1$ and $\alpha = 0.1$. The range of β is defined as $(0, \alpha]$. It can be observed that, adjusted protocols consistently achieve a lower MSE compared to original protocols, and as β increases, the MSE of adjusted protocols continuously decreases. This is due to an increase in the *inherent uncertainty* resulting in a decrease in *perturbation uncertainty*.

Figs. 5(c), 6(c), and 7(c) show the performance of the protocols under different privacy budget ε. We set $\alpha = 0.1$ and $\beta = 0.01$. The range of ε is defined as $[0.1, 3]$. It can be observed that, adjusted protocols consistently achieve a lower MSE compared to original protocols, and as ε increases, the MSE of adjusted protocols continuously decreases. This is due to a decrease in the *global uncertainty* resulting in a decrease in *perturbation uncertainty*.

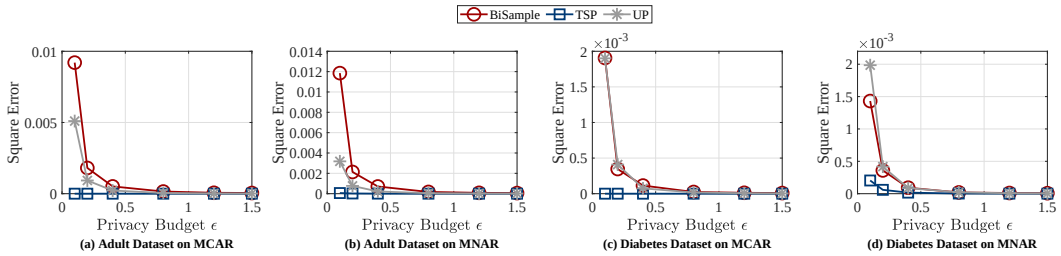


Fig. 8. Squared errors of nr for different methods.

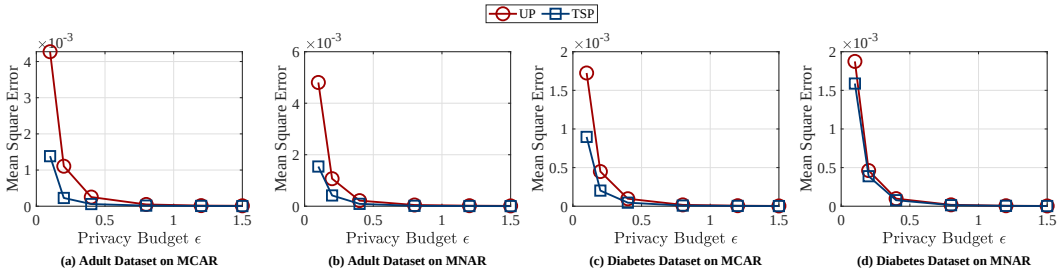


Fig. 9. MSE of f for *TSP*/*UP*.

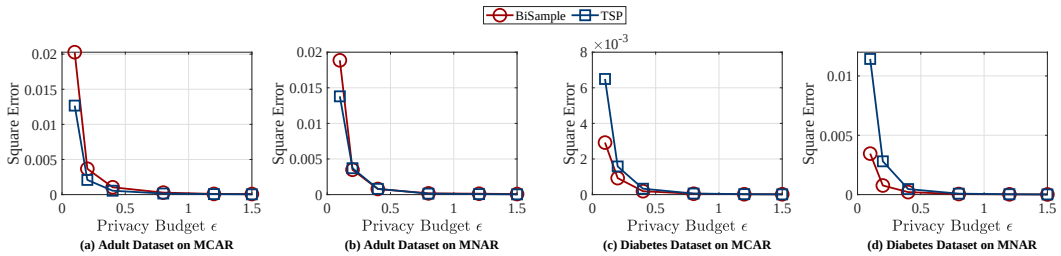


Fig. 10. Squared errors of m for *TSP*/*BiSample*.

Figs. 6(d), and 7(d) show the performance of *AGRR*/*GRR* and *AOUE*/*OUE* under different domain sizes of correlation attributes. We set $\alpha = 0.1$, $\beta = 0.01$, and $\varepsilon = 1$. It can be observed that, adjusted protocols consistently achieve a lower MSE compared to original protocols. In addition, from Fig. 6(d), we can observe that, as d increases, the MSE of *AGRR* and *GRR* increase. However, the increase in MSE of *AGRR* is relatively smaller, what is mainly attributed to AGRR's utilization of *inherent uncertainty*. From Fig. 7(d), we can observe that, as d increases, the MSE of *AOUE* and *OUE* remain almost unchanged, which is consistent with the fact that the variance of estimated frequency in *AOUE* and *OUE* is almost unaffected by d.

7.2.2 *Missing Data Collection.*

We first evaluate the performance of *TSP*, *UP*, and *Bisample* over real-world datasets, and then demonstrate the effectiveness of *TSP* and *UP* within a broader range of parameters.

Fig. 8 displays the squared errors in null value rate of *TSP* , *UP*, and *Bisample* over MCAR and MNAR missing datasets, across different privacy budgets. We can observe that our scheme consistently outperforms the *UP* and *Bisample*. The advantage is more pronounced when ε is lower.

Compared to MNAR, the advantage is more pronounced in the MCAR missing datasets. The reason lies in that, for MCAR missing datasets, the missing status does not reveal sensitive information about the relevant attributes, and thus there is no need to introduce perturbations to the missing status.

Fig. 9 displays the MSE in frequency estimation of *TSP* and *UP*. We can observe that our scheme consistently outperforms the *UP* scheme, with a more pronounced advantage seen in the MCAR missing datasets.

Fig. 10 displays the squared errors in mean estimation of *TSP* and *Bisample*. It can be observed that, our scheme's advantage is less in mean estimation. The reason lies in that our scheme faces the issue of noise accumulation when calculating the mean value.

Applicability and Distributional Assumptions. Our framework's reliance on attribute correlation contrasts with standard LDP's distribution-agnostic design. It is most effective when correlations are stable and learnable, e.g., from pilot studies, public data, or initial study phases. When correlations are unknown, standard LDP remains suitable; but when our assumptions hold, our method provides a substantial utility gain without compromising privacy.

7.3 Efficiency and Scalability Evaluation

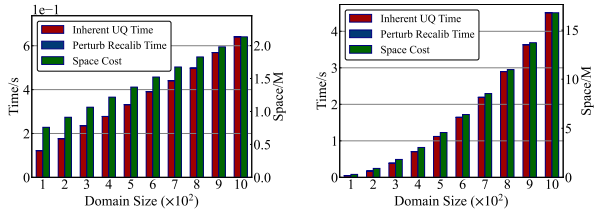
A core goal of our evaluation is to demonstrate that our schemes' additional operations incur only minimal overhead, thereby confirming its scalability. Our scheme introduces three modifications to standard LDP: (1) a one-time server-side computation for parameter recalibration; (2) broadcasting the recalibrated parameters (as one or two pairs of floats) to all users; and (3) a two-phase client-side perturbation for missing data collection. The broadcast overhead of (2) is negligible. Therefore, we focus our quantitative evaluation on the server overhead from (1) and the client overhead from (3).

Our scalability analysis uses synthetic data with varying attribute domain sizes ($d = 100$ to 1000). This approach directly controls the size of the core data structures (i.e., the conditional probability matrices), enabling us to isolate the effect of domain complexity and establish a distribution-agnostic scalability profile. Specifically, the experimental design is justified as follows:

- **Correlated Attribute Collection (ARR, AGRR, AOUE):** The server's extra computational and storage overhead is determined by the domain sizes of the collected non-sensitive attribute A ($|\Omega_A|$) and the correlated sensitive attribute S ($|\Omega_S|$). Thus, this overhead was measured by varying $|\Omega_A|$ and $|\Omega_S|$. (For ARR, $|\Omega_A|$ is fixed at 2).
- **Missing Data Collection (TSP):** With a fixed value space of 2 for the missing status, the server's overhead is determined by the domain size of the correlated attribute S ($|\Omega_S|$). Thus, the server's computational and storage overhead was tested by varying $|\Omega_S|$. The client-side perturbation time is determined by the domain size of the collected missing attribute A ($|\Omega_A|$). Thus, the client's computational cost was evaluated by varying $|\Omega_A|$.

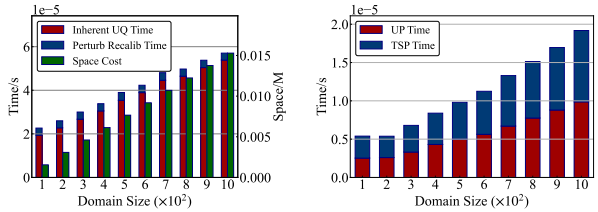
Due to space constraints, we present a subset of the most representative results. For the relevant attribute collection, we focus on the AOUE protocol, as it entails the most complex uncertainty parameter calculation among its counterparts (ARR, AGRR). This allows us to effectively demonstrate the scaling relationship between server-side overhead and the domain sizes ($|\Omega_A|$ and $|\Omega_S|$). For the missing data collection, we present the complete evaluation for the TSP scheme.

Fig. 11 presents the server-side performance evaluation of the AOUE protocol. Our measurements show a linear scaling of overhead with $|\Omega_A|$ and a quadratic scaling with $|\Omega_S|$. Notably, the uncertainty parameter calculation is identified as the primary computational bottleneck, with its cost far exceeding that of the perturbation step. Nevertheless, the overall overhead for both computation and storage is maintained at a low level, confirming the scheme's efficiency.



(a) Varying $|\Omega_A|$, and $|\Omega_S| = 200$ (b) Varying $|\Omega_S|$, and $|\Omega_A| = 200$

Fig. 11. Server-side performance of AOUE.



(a) Server side, varying $|\Omega_S|$ (b) Client side, varying $|\Omega_A|$

Fig. 12. Performance of the TSP.

Fig. 12 summarizes the performance of the TSP scheme. Our evaluation shows linear growth in server-side overhead with $|\Omega_S|$ and in client-side perturbation time with $|\Omega_A|$. The client-side time, however, is measured to be twice that of the unified scheme, a direct consequence of the requisite double perturbation for users with valid values for the missing attribute. Nevertheless, all measured overheads are maintained at a low level; the additional overhead introduced by our TSP scheme, when compared to the traditional protocol, remains minimal.

8 Conclusion

In this paper, we introduce an accuracy-enhancing framework by leveraging *inherent uncertainty* to address the indirect privacy risk problem and the leakage-pattern-driven hybrid privacy risk problem. To solve the problem of indirect privacy risk, we propose enhanced pure LDP protocols by (1) precisely quantifying the inherent uncertainty and (2) optimizing perturbation parameters based on the determined levels. To address the problem of the leakage-pattern-driven hybrid privacy risk, we propose the dual-phase perturbation scheme to maximize data utility while satisfying diverse privacy requirements across different values. These approaches are rigorously analyzed and extensively validated through theoretical and experimental studies, demonstrating the framework's correctness and effectiveness.

Acknowledgments

We thank the reviewers for their valuable comments which significantly improved this paper. This work was supported by the Heilongjiang Key R&D Program of China No. GA23A915, Key R&D Program of Shandong Province, China, No. 2024CXGC010114, National Natural Science Foundation of China, No. 62372268, No. 62572139, Shandong Provincial Natural Science Foundation, No. ZR2025MS1038, and Young Scholars Program of Shandong University.

References

[1] Jayadev Acharya, Kallista A. Bonawitz, Peter Kairouz, Daniel Ramage, and Ziteng Sun. 2020. Context Aware Local Differential Privacy. In *ICML*, Vol. 119. 52–62.

[2] Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *CCS*. 901–914.

[3] Raef Bassily and Adam Smith. 2015. Local, private, efficient protocols for succinct histograms. In *STOC*. 127–135.

[4] Albert Cheu, Adam D. Smith, Jonathan R. Ullman, David Zeber, and Maxim Zhilyaev. 2019. Distributed Differential Privacy via Shuffling. In *EUROCRYPT (Lecture Notes in Computer Science, Vol. 11476)*. 375–403.

[5] Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. 2018. Marginal Release Under Local Differential Privacy. In *SIGMOD*. 131–146.

[6] Badi Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. In *NeurIPS*. 3305–3314.

[7] Rong Du, Qingqing Ye, Yue Fu, and Haibo Hu. 2021. Collecting High-Dimensional and Correlation-Constrained Data with Local Differential Privacy. In *SECON*. 1–9.

[8] Flávio du Pin Calmon and Nadia Fawaz. 2012. Privacy against statistical inference. In *Allerton*. 1401–1408.

[9] Jiawei Duan, Qingqing Ye, and Haibo Hu. 2022. Utility Analysis and Enhancement of LDP Mechanisms in High-Dimensional Space. In *ICDE*. 407–419.

[10] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2018. Minimax optimal procedures for locally private estimation. *JASA* 113, 521 (2018), 182–201.

[11] Cynthia Dwork. 2006. Differential privacy. In *ICALP*. 1–12.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*. 265–284.

[13] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*. 1054–1067.

[14] Xiaolan Gu, Ming Li, Li Xiong, and Yang Cao. 2020. Providing Input-Discriminative Protection for Local Differential Privacy. In *ICDE*. 505–516.

[15] Mehmet Emre Gursoy, Acar Tamersoy, Stacey Truex, Wenqi Wei, and Ling Liu. 2021. Secure and Utility-Aware Data Collection with Condensed Local Differential Privacy. *TDSC* 18, 5 (2021), 2365–2378.

[16] Hsiang Hsu, Shahab Asoodeh, and Flávio P. Calmon. 2019. Information-Theoretic Privacy Watchdogs. In *ISIT*. 552–556.

[17] Bo Jiang, Ming Li, and Ravi Tandon. 2022. Local Information Privacy and Its Application to Privacy-Preserving Data Aggregation. *TDSC* 19, 3 (2022), 1918–1935.

[18] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. In *ICML*. 2436–2444.

[19] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2014. Extremal mechanisms for local differential privacy. In *NeurIPS*. 2879–2887.

[20] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. 2011. What Can We Learn Privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.

[21] Ninghui Li, Wahbeh H. Qardaji, and Dong Su. 2012. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *ASIACCS*. 32–33.

[22] Zitao Li, Tianhao Wang, Milan Lopuhaä-Zwakenberg, Ninghui Li, and Boris Skoric. 2020. Estimating Numerical Distributions under Local Differential Privacy. In *SIGMOD*. ACM, 621–635.

[23] Milan Lopuhaä-Zwakenberg. 2020. The Privacy Funnel from the viewpoint of Local Differential Privacy. *CoRR* abs/2002.01501 (2020).

[24] Milan Lopuhaä-Zwakenberg, Haochen Tong, and Boris Skoric. 2020. Data Sanitisation Protocols for the Privacy Funnel with Differential Privacy Guarantees. *CoRR* abs/2008.13151 (2020).

[25] Shubhankar Mohapatra, Jianqiao Zong, Florian Kerschbaum, and Xi He. 2024. Differentially Private Data Generation with Missing Data. *Proc. VLDB Endow.* 17, 8 (2024), 2022–2035.

[26] Takao Murakami and Yusuke Kawamoto. 2019. Utility-Optimized Local Differential Privacy Mechanisms for Distribution Estimation. In *USENIX*. 1877–1894.

[27] Qiuyu Qian, Qingqing Ye, Haibo Hu, Kai Huang, Tom Tak-Lam Chan, and Jin Li. 2023. Collaborative Sampling for Partial Multi-Dimensional Value Collection Under Local Differential Privacy. *TIFS* 18 (2023), 3948–3961.

[28] Xuebin Ren, Liang Shi, Weiren Yu, Shusen Yang, Cong Zhao, and Zongben Xu. 2022. LDP-IDS: Local Differential Privacy for Infinite Data Streams. In *SIGMOD*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). 1064–1077.

[29] Lin Sun, Xiaojun Ye, Jun Zhao, Chenhui Lu, and Mengmeng Yang. 2020. BiSample: Bidirectional Sampling for Handling Missing Data with Local Differential Privacy. In *DASFAA*, Vol. 12112. 88–104.

[30] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019. Collecting and Analyzing Multidimensional Data with Local Differential Privacy. In *ICDE*. 638–649.

[31] Shaowei Wang, Yiwen Nie, Pengzhan Wang, Hongli Xu, Wei Yang, and Liusheng Huang. 2017. Local private ordinal data distribution estimation. In *INFOCOM*. IEEE, 1–9.
[32] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally differentially private protocols for frequency estimation. In *USENIX Security*. 729–745.
[33] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. 2017. Locally Differentially Private Protocols for Frequency Estimation. In *USENIX*. 729–745.
[34] Tianhao Wang, Milan Lopuhaä-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. 2020. Locally Differentially Private Frequency Estimation with Consistency. In *NDSS*.
[35] Stanley L. Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.
[36] Emre Yilmaz, Tianxi Ji, Erman Ayday, and Pan Li. 2022. Genomic Data Sharing under Dependent Local Differential Privacy. In *CODASPY*. 77–88.
[37] Mohammad Amin Zarrabian, Ni Ding, and Parastoo Sadeghi. 2023. On the Lift, Related Privacy Measures, and Applications to Privacy-Utility Trade-Offs. *Entropy* 25, 4 (2023), 679.
[38] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, and Jiming Chen. 2018. CALM: Consistent Adaptive Local Marginal for Marginal Release under Local Differential Privacy. In *CCS*. 212–229.

Received July 2025; revised October 2025; accepted November 2025