Addressing Sensitivity Distinction in Local Differential Privacy: A General Utility-Optimized Framework

Xingyu He[1], Youwen Zhu[1,*], Rongke Liu[1], Gaoning Pan[2], Changyu Dong[3]

[1]Nanjing University of Aeronautics and Astronautics
[2]Hangzhou Dianzi University, [3]Guangzhou University
[*]Corresponding author: *zhuyw@nuaa.edu.cn*

Abstract

Local Differential Privacy (LDP) is widely employed to address privacy concerns in data collection. Nevertheless, the LDP model ignores the sensitivity distinction, as it regards all personal data equally sensitive, leading to excessive obfuscation and the loss of utility. Utility-optimized LDP (ULDP) aims to mitigate this issue. However, existing ULDP mechanisms address sensitivity distinction in only a limited subset of LDP mechanisms. To systematically address sensitivity distinction in the LDP model, we propose the General LDP-to-ULDP Transformation Framework. This framework can convert any LDP mechanism into its corresponding ULDP mechanism while preserving key properties such as order-optimality and unbiased estimation. Then, we present the pure ULDP framework, which generalizes a class of ULDP mechanisms with strong performance guarantees. We develop a universal aggregation and utility analysis method applicable to all pure ULDP mechanisms, facilitating the analysis, comparison, and optimization of different ULDP mechanisms. After that, we transform three widely-used LDP mechanisms into their ULDP counterparts (uSS, uUE and uLH). We theoretically demonstrate that our proposed mechanisms exceed existing ULDP mechanisms in data utility and communication costs. Specifically, our uSS, uUE and uLH match the minimax risk lower bound within the ULDP framework. We also identify the optimal mechanism for various usage scenarios. Finally, we conduct experiments on both real and synthetic datasets, showing that uUE and uLH achieve the lowest Mean Squared Error (MSE) when size of sensitive dataset is large, and uSS consistently achieves the lowest MSE.

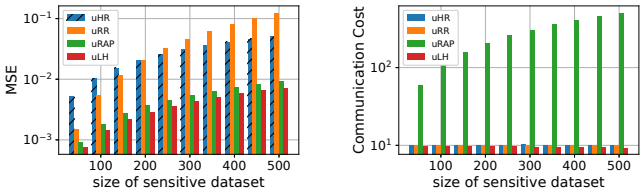
1 Introduction

Differential Privacy (DP) [1, 2] provides low computational overhead and robustness against attackers with arbitrary background knowledge, and has become the *de facto* standard in privacy-preserving data collection scenarios. However, it requires a trusted third party for data aggregation, which is impractical in many application scenarios. To address this issue, Local Differential Privacy (LDP) [3] was proposed, which inherits the advantages of DP while eliminating the need for a trusted third party. In the local setting, the server is assumed to be untrusted, and each user locally perturbs their private data. The server then collects the perturbed data from each user and performs statistical estimation. LDP is suitable for distributed data collection and has been widely adopted by industry. For example, Google [4], Microsoft [5], and Apple [6] have used LDP in their applications to collect user information while preserving privacy.

One of the most important problems in LDP is frequency estimation for categorical data. Improving this fundamental task can not only yield more accurate frequency estimates, but also facilitate advancements in more complex tasks that rely on it, such as heavy-hitter identification [7] and relation mining [8]. Although various LDP frequency estimation mechanisms have been proposed, the LDP model has inherent limitations that hinder further improvements in data utility.

LDP assumes that all private data are equally sensitive, which is unrealistic in practical applications and leaves significant room for improving data utility. For example, when counting diseases, conditions like "AIDS" and "cancer" require stricter privacy protection compared to conditions like "colds", as the sensitivity of the former is much higher. Treating all data as equally sensitive leads to what we refer to as "**a lack of sensitivity distinction**", inevitably resulting in excessive obfuscation, which in turn diminishes data utility.

Private data naturally exhibit differences in sensitivity, and differentiating between sensitive and non-sensitive data enables focused protection where it is most needed. In practice, users (although not privacy experts) often have intuitive judgments. For example, individuals are reluctant to share precise home addresses online but are more comfortable disclosing their workplace; similarly, detailed medical records are protected carefully, while step counts or workout summaries are frequently shared on social media. Motivated by this observation, recent works have introduced a privacy notion called Utility-optimized LDP (ULDP) [9] ([10] and [11]



(a) Data utility (b) Communication cost

Figure 1: Evaluation of data utility and communication cost in existing ULDP mechanisms (uRR, uRAP and uHR) vs. one of our proposed mechanisms (uLH) on large sensitive dataset. Due to limitations in theoretical analysis, uHR in (a) represents only the upper bound of MSE, with a significant gap from the actual MSE.

independently proposed similar models). The ULDP framework categorizes data into sensitive and non-sensitive groups, applying LDP-level privacy guarantees only to the sensitive subset, thereby improving overall data utility. Importantly, ULDP does not rely on a universal definition of sensitive data; instead, it accommodates user-specific sensitivity through a personalization framework [9].

Although ULDP shows great promise, its full potential has yet to be reached. To the best of our knowledge, there are currently only three ULDP mechanisms: utility-optimized Randomized Response (uRR) [9], utility-optimized RAPPOR (uRAP) [9], and utility-optimized Hadamard Response (uHR) [11]. The utility of existing ULDP mechanisms is often suboptimal in various scenarios. For example, as illustrated in Fig. 1, when the size of the sensitive dataset increases, the data utility of the uRR mechanism decreases rapidly, while the communication cost of the uRAP mechanism increases significantly. Furthermore, within the medium privacy regime ($1 < \varepsilon < \log s$, where s is the size of sensitive dataset), the uRR and uRAP mechanisms do not achieve order-optimality [9], leading to the introduction of unnecessary noise. Although the uHR mechanism achieves a low communication cost, its design inherently prevents it from providing an accurate theoretical MSE, making it challenging to determine its application scenarios. These limitations hinder their broader applicability and efficiency of these mechanisms in real-world data collection tasks.

Thus, the issue of sensitivity distinction in the LDP model has yet to be systematically addressed. This is primarily because current ULDP-related research focuses on developing ULDP versions for individual LDP mechanisms, thereby tailoring them to specific scenarios. The limitations of existing ULDP mechanisms highlight that patchwork solutions alone cannot effectively resolve the problem. A truly comprehensive solution must capitalize on the close relationship between the LDP and ULDP models to establish a universal transformation framework, allowing any LDP mechanism to be converted into a ULDP mechanism. It is essential to build a bridge between the LDP and ULDP models. This insight has motivated our research.

To systematically address the issue of sensitivity distinction in the LDP model, we must overcome two major challenges: generalizability and high data utility. For the first challenge, both LDP and ULDP mechanisms exhibit significant differences in their perturbation and aggregation. The variations in perturbation probabilities and privacy budget allocation across different mechanisms make it difficult to establish a unified mathematical notation for consistent representation, let alone design a generalized perturbation and aggregation framework based on it. For the second challenge, when transforming an LDP mechanism into its ULDP version, it is crucial to carefully control the introduction of noise to avoid unnecessary utility loss. Furthermore, maintaining key theoretical properties of the original LDP mechanism, such as order optimality and unbiased estimation, throughout the transformation process presents an additional layer of complexity.

In this paper, we propose a General LDP-to-ULDP Transformation Framework (GLUTF) that can transform any LDP mechanism into a ULDP mechanism. The core concept of GLUTF is to utilize one existing LDP mechanism as a basic building block, allowing users to apply it differently depending on the sensitivity of their private data. GLUTF is highly adaptable and can be applied to any LDP mechanism.

Furthermore, we propose the pure ULDP framework, which generalizes a class of ULDP mechanisms with desirable performance. We develop a simple and general aggregation and utility analysis method applicable to all pure ULDP mechanisms, ensuring unbiased estimation and providing an accurate theoretical mean squared error (MSE). This framework enables the precise evaluation and comparison of different mechanisms in terms of data utility and facilitates the optimization of the mechanism based on theoretical MSE. Through utility analysis, we found that if an LDP mechanism is inherently order-optimal or provides unbiased estimates, these desirable properties are preserved after transformation to a ULDP mechanism via GLUTF.

Finally, based on three widely used and efficient LDP mechanisms, we propose three new ULDP mechanisms: utility-optimized Subset Selection (uSS), utility-optimized Unary Encoding (uUE), and utility-optimized Local Hashing (uLH). We demonstrate that these mechanisms achieve order-optimality within commonly used privacy regime ($0 < \varepsilon < \log s$), and the uLH mechanism can simultaneously provide high data utility and low communication cost when the size of sensitive dataset is large.

Our contributions are summarized as follows:

- We propose the GLUTF framework that enables the conversion of any LDP mechanism into its corresponding ULDP mechanism. We also demonstrate that GLUTF preserves the order-optimality and unbiased estimation properties of the underlying LDP protocols.

- We propose the pure ULDP framework and develop a general aggregation and utility analysis method applicable to all pure ULDP mechanisms. This framework enables the analysis, comparison, and optimization of different ULDP mechanisms. We also demonstrate how to extend these methods to non-pure ULDP mechanisms.
- We develop three new ULDP mechanisms: uSS, uUE and uLH. The MSE of each mechanism is $O(\frac{s}{n\varepsilon^2})$ $(0<\varepsilon<1)$ or $O(\frac{se^{\varepsilon}}{n(e^{\varepsilon}-1)^2})$ $(1<\varepsilon<\log s)$, indicating that these mechanisms have achieved order-optimality. The communication cost of uLH is $O(\log(e^{\varepsilon}+d-s))$, which is better than that of all existing ULDP mechanisms (when $\varepsilon<\ln s$).
- We conduct systematic experiments on real and simulated datasets. The experimental results validate the design and analysis of the GLUTF and the pure ULDP framework, demonstrating their advantages and effectiveness.

2 Preliminaries

2.1 Problem Definition and Notations

In this paper, we focus on frequency estimation for categorical data. Our system model involves one data server and n users. We assume each user possesses a single categorical value, and the server aims to determine the proportion of users with a specific private value. To ensure privacy, users locally encode and perturb their data before sending it to the server.

Formally, the problem is defined as follows: there are n users, each possessing a **private data** $x \in X$, where $X = \{1,2,\ldots,d\}$. The frequency distribution of the private data is represented by the vector $\boldsymbol{c} = (c_1, c_2, \ldots, c_d)$, where c_x represents the frequency of private data x. Each user encodes their private data x into **encoded data** $\dot{x}$, then applies a perturbation mechanism, producing the **perturbed data** $y \in Y$. For simplicity, we denote the combined encoding and perturbation processes as $\mathcal{A}$, defined as $y = \mathcal{A}(x)$. In addition, there is a server that collects perturbed data from each user and performs statistical estimation to obtain $\hat{\boldsymbol{c}} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_d)$, an estimate of $\boldsymbol{c}$.

Based on data sensitivity, the private data set X is divided into two subsets: the sensitive data set X_S and the non-sensitive data set X_N, where $|X_S| = s$, $|X_N| = d - s$ and $X_N = X \setminus X_S$. The total frequency of all non-sensitive data is expressed as $\theta = \sum_{x \in X_N} c_x$. The perturbed data set Y is divided into the protected data set Y_P and the invertible data set Y_I, where $Y_P = \{y | x \in X_S, Pr[y = \mathcal{A}(x)] > 0\}$ and $Y_I = Y \setminus Y_P$.

Threat Model. As a convention, we assume the semi-honest security model. Our focus is on protecting individual privacy, without considering robustness against user misbehavior or data poisoning.

2.2 Local Differential Privacy

Local Differential Privacy (LDP) [3], a privacy model that allows data collection without a trusted third party, is defined as follows:

Definition 1 (ε-LDP [3]). *Given $\varepsilon > 0$. A randomized algorithm $\mathcal{A}: X \to Y$ satisfies ε-LDP if and only if for any $x_1, x_2 \in X$ and any $y \in Y$, we have:*

$$\Pr[\mathcal{A}(x_1) = y] \leq e^{\varepsilon} \Pr[\mathcal{A}(x_2) = y]. \tag{1}$$

In Definition 1, ε is referred to as the **privacy budget**. This parameter quantifies how closely the perturbed outcomes of two different pieces of private data resemble each other, thus serving as a measure of the strength of privacy protection. A larger privacy budget leads to reduced privacy protection but enhances data utility.

2.3 Pure LDP

Wang et al. [12] proposed the notion of pure LDP protocols. To be pure, an LDP protocol requires a "support" function, denoted as $\mathrm{Supp}(\cdot)$, which maps each output y to a set of input values it supports. Pure LDP is defined as follows:

Definition 2 (Pure LDP [12]). *Given $\mathrm{Supp}(\cdot)$. An LDP mechanism $\mathcal{A}: X \to Y$ is pure if and only if there exist two probability values $p^* > q^*$ such that for all $x_1 \in X$,*

$$\Pr[\mathcal{A}(x_1) \in \{y | x_1 \in \mathrm{Supp}(y)\}] = p^*, \tag{2}$$

$$\forall_{x_2 \neq x_1} \Pr[\mathcal{A}(x_2) \in \{y | x_1 \in \mathrm{Supp}(y)\}] = q^*, \tag{3}$$

where p^* and q^* are called pure probabilities, and $\{y | x_1 \in \mathrm{Supp}(y)\}$ is referred to as the support set of x_1.

For input x, its "support set" is a subset of the output space where x is more likely to fall (with probability p^*) than other input (with probability q^*). Obviously, pure LDP is a strict subset of LDP. Most existing LDP protocols for frequency estimation are pure, including GRR [13], basic RAPPOR [4], SS [14, 15], OUE [12], OLH [12], and Wheel [16].

2.4 Utility-Optimized LDP

The Utility-optimized LDP (ULDP) model [9] is a variant of LDP. In the ULDP model, private data is divided into two types: sensitive data and non-sensitive data. This novel privacy model provides the same level of protection as LDP only for sensitive data, hence significantly enhancing data utility. ULDP is defined as follows:

Definition 3 ((X_S, Y_P, ε)-ULDP [9]). *Given $\varepsilon > 0$, $X_S \subset X$, $X_N = X \setminus X_S$, $Y_P \subset Y$ and $Y_I = Y \setminus Y_P$, an randomized algorithm $\mathcal{A}: X \to Y$ satisfies (X_S, Y_P, ε)-ULDP if and only if it satisfies the following properties:*

1. *For any $y \in Y_I$, there exists an $x \in X_N$ such that*

$$\Pr[\mathcal{A}(x) = y] > 0 \text{ and } \Pr[\mathcal{A}(x') = y] = 0 \text{ for any } x' \neq x, \tag{4}$$

2. *For any $x, x' \in X$, any $y \in Y_p$*

$$\Pr[\mathcal{A}(x) = y] \leq e^{\varepsilon} \Pr[\mathcal{A}(x') = y]. \tag{5}$$

(X_S, Y_P, ε)-ULDP provides a privacy guarantee equivalent to ε-LDP for any sensitive data $x \in X_s$ (with regard to a restricted output domain Y_P). For non-sensitive data $x \in X_N$, no privacy guarantee is provided, thereby enabling the server to obtain a more accurate estimation. To the best of our knowledge, only three ULDP mechanisms exist: uRR [9], uRAP [9] and uHR [11].

2.5 Utility Metrics

We assess the data utility of the mechanism using Mean Squared Error (MSE), a utility metric widely used in prior studies. MSE is defined as follows:

$$MSE[\hat{\boldsymbol{c}}] = E[\| \hat{\boldsymbol{c}} - \boldsymbol{c} \|_2^2]. \tag{6}$$

A smaller MSE indicates a closer approximation to the true values. When $\hat{\boldsymbol{c}}$ is an unbiased estimate of the true frequency $\boldsymbol{c}$, the MSE corresponds to the average of the variances:

$$MSE[\hat{\boldsymbol{c}}] = \sum_{i=1}^{d} Var[\hat{c}_i]. \tag{7}$$

Lower bounds on the l_2 losses. Ye et al. [15] showed that the lower bounds on the l_2 losses (minimax rates) of any ε-LDP mechanism is $\Theta(\frac{d}{n\varepsilon^2})$ (when $\varepsilon \in (0,1)$) and $\Theta(\frac{de^{\varepsilon}}{n(e^{\varepsilon}-1)^2})$ (when $\varepsilon \in (1, \log d)$). By directly applying these bounds to X_S and Y_P, the lower bounds on the l_2 losses of any ULDP mechanisms can be derived as $\Theta(\frac{s}{n\varepsilon^2})$ (when $\varepsilon \in (0,1)$) and $\Theta(\frac{se^{\varepsilon}}{n(e^{\varepsilon}-1)^2})$ (when $\varepsilon \in (1, \log s)$). In Section 5.4, we will show that when $\varepsilon < \log s$, our ULDP mechanism match the minimax lower bound of ULDP model. This property is commonly referred to as *order-optimality*, a standard notion in LDP indicating near-optimal utility, which is defined as follows:

Definition 4 (Order-optimal)**.** *A mechanism is order-optimal if its error achieves the minimax lower bound up to constant factors.*

We also consider the communication cost of the mechanism, which is defined as follows:

Definition 5 (Communication cost)**.** *The communication cost of a mechanism is defined as the minimum number of bits needed to uniquely represent an output. For an algorithm $\mathcal{A} : X \to Y$, its communication cost scales as $O(\log |Y|)$.*

In this definition, we consider only the perturbed data y sent from the user to the server, excluding any overhead from prior sharing of sensitive/non-sensitive sets or other coordination information.

Algorithm 1 Perturbation of GLUTF

Input: Sensitive data set X_S, non-sensitive data set X_N, pure LDP mechanism $\mathcal{A}_{LDP}$, pure probabilities p^*_{LDP} and q^*_{LDP}, private data x, perturbation probabilities z and f
Output: perturbed data y

```
if x ∈ X_S then
    y = A_LDP(x)
else
    if UniformRandom(0.0, 1.0) < f then
        Uniformly randomly select an element x' from X_S
        if UniformRandom(0.0, 1.0) < z then
            y = ⟨A_LDP(x'), x⟩
        else
            y = A_LDP(x')
        end if
    else
        y = x
    end if
end if
```

3 Generalized LDP-to-ULDP Transformation Framework

To improve the data utility of LDP mechanisms and develop more efficient ULDP mechanisms, we propose the Generalized LDP-to-ULDP Transformation Framework (GLUTF) that can convert any LDP mechanisms into their corresponding ULDP counterparts. Although pure and non-pure LDP differ significantly, we find that they share similar structures and steps in the process of converting to ULDP mechanisms.

GLUTF's intuition is simple: it utilizes an existing LDP mechanism $\mathcal{A}_{LDP}$ as its basic building block and applies it to sensitive and non-sensitive data differently to achieve ULDP conversion. The overview of GLUTF is shown in Fig. 2.

Detailed steps of GLUTF are shown in Algorithm 1. Given sensitive data set X_S and any LDP mechanism $\mathcal{A}_{LDP} : X_S \to Y_{LDP}$, private data with different sensitivities will be perturbed in different ways. For any sensitive data $x \in X_S$, it is directly used as the input of $\mathcal{A}_{LDP}$, producing the perturbed data $y = \mathcal{A}_{LDP}(x)$. For any non-sensitive data $x \in X_N$, it remains unaltered with a probability of $1 - f$, and it maps to any sensitive data with a probability of $\frac{f}{s}$, where $s = |X_S|$.

If x is retained as is, it is directly output. In case where x is mapped to sensitive data x', the output is $\mathcal{A}_{LDP}(x')$, and x is also output simultaneously with probability z. In summary, the probability of $x \in X_N$ producing output y is as follows:

$$\Pr[y = i] = \begin{cases} f(1-z), & \text{if } i = \mathcal{A}_{LDP}(x') \\ fz, & \text{if } i = \langle \mathcal{A}_{LDP}(x'), x \rangle \\ 1 - f, & \text{if } i = x, \end{cases} \tag{8}$$

where $\langle \mathcal{A}_{LDP}(x'), x \rangle$ represents the simultaneous output of $\mathcal{A}_{LDP}(x')$ and x.

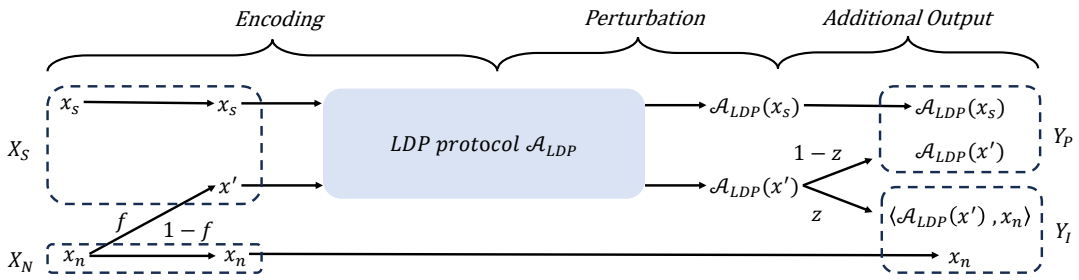


Figure 2: Overview of GLUTF. Dashed boxes illustrate input/output partitions; arrows indicate data perturbation pathways.

f Selection: Pure vs. Non-pure. Although the pure LDP and non-pure LDP mechanisms follow the same transformation process in GLUTF, key differences remain in the choice of the parameter f. In the pure LDP mechanism, any private data maps to its own support set with probability p^*, and to any other private data's support set with probability q^*. This strict probability drives a simple and general aggregation method and utility analysis method for pure LDP mechanism. To preserve this property, we set f as follows:

$$f = \frac{sq^*_{LDP}}{p^*_{LDP} + (s-1)q^*_{LDP}}. \tag{9}$$

With this parameter setting, the ULDP mechanisms converted by pure LDP mechanisms retain a general aggregation method and a utility analysis method, which will be explored in Section 4. In contrast, the non-pure LDP mechanisms lack restrictions on perturbation probability, making it challenging to identify commonalities. Therefore, we can only rely on the basic parameters defined in the LDP model to set f:

$$f = \frac{s}{e^\varepsilon + s - 1} \tag{10}$$

We also discussed how to aggregate and analyze this type of ULDP mechanisms in Section 4.

The domain of the perturbed data is comprised of three parts: Y_{LDP}, X_N and $\{\langle y_1, y_2\rangle | y_1 \in Y_{LDP}, y_2 \in X_N\}$, where Y_{LDP} is the output domain of $\mathcal{A}_{LDP}$. To adapt the ULDP structure, we divide the perturbed data into:

$$Y_P = Y_{LDP}, \tag{11}$$

$$Y_I = X_N \cup \{\langle y_1, y_2\rangle | y_1 \in Y_{LDP}, y_2 \in X_N\}. \tag{12}$$

After presenting the core ideas and parameter selection of GLUTF, we will next prove that the protocol generated by this framework indeed satisfies the definition of ULDP.

Theorem 1. *The mechanism $\mathcal{A}$ we construct based on any LDP mechanism satisfies the ULDP model when $z \le 1 - \frac{1}{e^\varepsilon}\max\{\frac{p_{im}}{\sum_{t=1}^{s}\frac{f}{s}p_{tm}}\}$, where p_{im} is the probability that x_i maps to y_m in the LDP mechanism.*

Proof. For any $y \in Y_I$, it necessarily contains an element $x \in X_N$. Evidently, the output y is possible only if x is the input. Furthermore, any $x' \neq x$ cannot map to y. Thus, the mechanism $\mathcal{A}$ satisfies the first property outlined in Definition 3.

The intuition behind the following proof is to perform a case analysis, verifying whether $\mathcal{A}$ satisfies the second property outlined in Definition 3.

For any $x_i, x_j \in X_S$ and $y_m \in Y_P$, it is evident that their perturbation processes are identical to those of a LDP mechanism. Consequently, they necessarily satisfy $\Pr[\mathcal{A}(x_i) = y_m] \le e^\varepsilon \Pr[\mathcal{A}(x_j) = y_m]$. For any $x_i, x_j \in X_N$ and $y_m \in Y_P$, it is manifest that $\Pr[\mathcal{A}(x_i) = y_m] = \Pr[\mathcal{A}(x_j) = y_m]$ holds.

For any $x_i \in X_S$, $x_j \in X_N$ and $y_m \in Y_P$, we have:

$$\frac{\Pr[\mathcal{A}(x_j) = y_m]}{\Pr[\mathcal{A}(x_i) = y_m]} = \frac{(1-z)\sum_{t=1}^{s}\frac{f}{s}p_{tm}}{p_{im}}, \tag{13}$$

where p_{im} denotes the probability that x_i maps to y_m in the LDP mechanism $\mathcal{A}_{LDP}$. From Definition 1, it becomes apparent that $p_{tm} \le e^\varepsilon p_{im}$, so we have:

$$\frac{\Pr[\mathcal{A}(x_j) = y_m]}{\Pr[\mathcal{A}(x_i) = y_m]} \le (1-z) f e^\varepsilon \le e^\varepsilon. \tag{14}$$

Let's consider another case:

$$\frac{\Pr[\mathcal{A}(x_i) = y_m]}{\Pr[\mathcal{A}(x_j) = y_m]} = \frac{p_{im}}{(1-z)\sum_{t=1}^{s}\frac{f}{s}p_{tm}}. \tag{15}$$

We also know that $p_{im} \le e^\varepsilon p_{tm}$, then we get:

$$\frac{\Pr[\mathcal{A}(x_i) = y]}{\Pr[\mathcal{A}(x_j) = y]} \le \frac{se^\varepsilon}{(1-z)f(e^\varepsilon + s - 1)}. \tag{16}$$

Regardless of whether GLUTF uses a pure (Eq. (9)) or non-pure (Eq. (10)) LDP protocol, we can obtain $\frac{\Pr[\mathcal{A}(x_i)=y]}{\Pr[\mathcal{A}(x_j)=y]} \le e^\varepsilon$ if $z = 0$. However, it is important to note that in many cases, setting z to 0 results in a waste of the privacy budget. Specifically, in such cases, $\frac{\Pr[\mathcal{A}(x_i)=y]}{\Pr[\mathcal{A}(x_j)=y]} = e^{\varepsilon'} < e^\varepsilon$, where $\varepsilon' < \varepsilon$. Therefore, to maximize the data utility while satisfying the ULDP, we generally aim to increase the value of z. The optimal z can be determined as $1 - \frac{1}{e^\varepsilon}\max\{\frac{p_{im}}{\sum_{t=1}^{s}\frac{f}{s}p_{tm}}\}$. In summary, $\mathcal{A}$ satisfies the second property outlined in Definition 3. □

In GLUTF, the parameter z controls the probability of disclosing the true value of non-sensitive data after it has been

disguised as sensitive data. The core purpose of this parameter is to leak as much information as possible while still meeting privacy requirements. Also within GLUTF, different LDP protocols have varying privacy guarantees after undergoing the same transformation process. The introduction of z enables fine-grained adjustments to different protocols. This strategy not only enhances the data utility of the mechanism but also improves GLUTF's compatibility with different LDP mechanisms.

4 Pure ULDP Framework

We propose the pure ULDP framework, summarizing a class of concise and efficient ULDP mechanisms. In this section, we provide simple, generic aggregation and utility analysis methods for both pure and non-pure ULDP mechanisms. Furthermore, we theoretically prove that the "pure" property is preserved during the GLUTF transformation, enabling us to swiftly apply the aggregation and utility analysis methods to the ULDP mechanisms constructed by GLUTF.

We identified a class of ULDP mechanisms with advantageous properties. Specifically, each private data corresponds to a perturbed data set, referred to as the support set. All private data within the same sensitivity are mapped to their own support set with the same probability, and all other private data are mapped to this set with another uniform probability. To define the support set, we introduced the function Supp. $\text{Supp}(y)$ represents all private data supported by the perturbation data y, and $\{y|x \in \text{Supp}(y)\}$ represents the support set of the private data x. We propose the pure ULDP framework to encapsulate the above mechanisms, which is defined as follows:

Definition 6 (pure ULDP)**.** *Given the* Supp *function, a ULDP mechanism $\mathcal{A}$ is pure if and only if there exist three probability values $p^* > q^*$ and z^* that satisfy the following properties:*

$$\forall_{x_1 \in X_S} \Pr[\mathcal{A}(x_1) \in \{y|x_1 \in \text{Supp}(y)\}] = p^*, \tag{17}$$

$$\forall_{x_2 \neq x_1 \in X} \Pr[\mathcal{A}(x_2) \in \{y|x_1 \in \text{Supp}(y)\}] = q^*, \tag{18}$$

$$\forall_{x_3 \in X_N} \Pr[\mathcal{A}(x_3) \in \{y|x_3 \in \text{Supp}(y)\}] = z^*, \tag{19}$$

$$\forall_{x_4 \neq x_3 \in X} \Pr[\mathcal{A}(x_4) \in \{y|x_3 \in \text{Supp}(y)\}] = 0, \tag{20}$$

where p^*, q^* and z^* are called pure probabilities.

Using p^*, q^* and z^* to define perturbation probabilities, we can use the same symbols to represent different encoding and perturbation processes, which is a necessary prerequisite for general theoretical analysis.

For any pure ULDP mechanism, the server can estimate the frequency of private data x as following:

$$\hat{c}_x = \begin{cases} \dfrac{\sum_{i=1}^{n} \mathbb{1}_{support(y_i)}(x) - nq^*}{n(p^* - q^*)}, & \text{if } x \in X_S \\ \dfrac{\sum_{i=1}^{n} \mathbb{1}_{support(y_i)}(x)}{nz^*}, & \text{if } x \in X_N, \end{cases} \tag{21}$$

where y_i is the perturbed data of user i, and $\mathbb{1}_{support(y)}(x)$ is a indicator function. The function determines whether a perturbed data y supports the private data x, defined as follows:

$$\mathbb{1}_{support(y)}(x) = \begin{cases} 1, & \text{if } x \in \text{Supp}(y) \\ 0, & \text{if } x \notin \text{Supp}(y). \end{cases} \tag{22}$$

Theorem 2. *For any pure ULDP mechanism, the estimated frequency $\hat{c}_x$ in Eq. (21) is an unbiased estimate of the true frequency c_x.*

Proof. For $x \in X_S$, we observe that $\sum_{j=1}^{n} \mathbb{1}_{support(y_i)}(x)$ can be viewed as the sum of two binomial distributions, $B(nc_x, p^*)$ and $B(n(1-c_x), q^*)$. Thus, we have

$$E[\hat{c}_x] = \frac{nc_x p^* + n(1-c_x)q^* - nq^*}{n(p^* - q^*)} = c_x. \tag{23}$$

For $x \in X_N$, $\sum_{j=1}^{n} \mathbb{1}_{support(y_i)}(x)$ equates to the outcome of a single binomial distribution $B(nc_x, z^*)$. Thus, we have

$$E[\hat{c}_x] = \frac{nc_x z^*}{nz^*} = c_x. \tag{24}$$

In summary, $\hat{c}_x$ is an unbiased estimate of c_x. □

Having introduced a unified frequency estimation approach within the pure ULDP framework, the subsequent question arises regarding the accuracy of these estimations. In the following convention, we utilize MSE as the utility metric. As demonstrated in Section 2.5, MSE is equivalent to the sum of variance for unbiased estimations. Therefore, we first calculate the variance of $\hat{c}_x$.

Theorem 3. *For any pure ULDP mechanism, the variance of the estimation $\hat{c}_x$ in Eq. (21) is:*

$$Var[\hat{c}_x] = \begin{cases} c_x \dfrac{1 - p^* - q^*}{n(p^* - q^*)} + \dfrac{q^*(1-q^*)}{n(p^* - q^*)^2}, & \text{if } x \in X_S \\ c_x \dfrac{1 - z^*}{nz^*}, & \text{if } x \in X_N. \end{cases} \tag{25}$$

Proof. The variance of the estimated value can be calculated by utilizing the characteristics of binomial distribution. For $x \in X_S$, we have:

$$\begin{aligned} Var[\hat{c}_x] &= Var[\frac{\sum_{i=1}^{n} \mathbb{1}_{support(y_i)}(x) - nq^*}{n(p^* - q^*)}] \\ &= \frac{Var[\sum_{i=1}^{n} \mathbb{1}_{support(y_i)}(x)]}{n^2(p^* - q^*)^2} \\ &= \frac{nc_x p^*(1-p^*) + n(1-c_x)q^*(1-q^*)}{n^2(p-q)^2} \\ &= c_x \frac{1 - p^* - q^*}{n(p^* - q^*)} + \frac{q^*(1-q^*)}{n(p^* - q^*)^2}. \end{aligned} \tag{26}$$

For $x \in X_N$, we have:

$$Var[\hat{c}_x] = Var[\frac{\sum_{i=1}^{n} \mathbb{1}_{support(y_i)}(x)}{nz^*}] = c_x \frac{1-z^*}{nz^*}. \quad (27)$$

□

Combining Eq. (7) and (25), we can directly obtain the MSE of $\hat{\boldsymbol{c}} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_d)$.

Theorem 4. *For any pure ULDP mechanism, the MSE of the estimation $\hat{\boldsymbol{c}}$ is:*

$$\frac{1}{n}((1-\theta)\frac{1-p^*-q^*}{(p^*-q^*)} + s\frac{q^*(1-q^*)}{(p^*-q^*)^2} + \theta\frac{1-z^*}{z^*}). \quad (28)$$

We proposed general aggregation and utility analysis methods for pure ULDP mechanisms, offering powerful tools for analyzing and comparing these mechanisms. A natural question arises: can the ULDP mechanisms generated by GLUTF in Section 3 benefit from these tools? Fortunately, we found that the "pure" property is preserved during the GLUTF conversion. Specifically, when we use GLUTF to convert any pure LDP protocol, the resulting ULDP mechanism remains pure. The Supp function for this pure ULDP is as follows:

$$\text{Supp}(y) = \begin{cases} \text{Supp}_{LDP}(y), & \text{if } y \in Y_P \\ \{y\}, & \text{if } y \in X_N \\ \text{Supp}_{LDP}(y_1) \cup \{y_2\}, & \text{if } y = \langle y_1, y_2 \rangle, \end{cases} \quad (29)$$

where Supp_{LDP} is the support function of the pure LDP mechanism and $\langle y_1, y_2 \rangle$ represents the simultaneous output of y_1 and y_2.

Theorem 5. *Given the* Supp *function as defined in Eq. (29), the ULDP mechanism $\mathcal{A}$ generated by GLUTF from a pure LDP mechanism must be pure. Specifically, $p^* = p^*_{LDP}$, $q^* = q^*_{LDP}$ and $z^* = (1-f) + fz$, where p^*_{LDP} and q^*_{LDP} are the pure probabilities of the pure LDP mechanism.*

Proof. For any private data $x \in X_S$, $\{y | x \in \text{Supp}(y)\} = \{y | x \in \text{Supp}_{LDP}(y)\} \cup \{\langle y_1, y_2 \rangle | x \in \text{Supp}_{LDP}(y_1)\}$. For any private data $x \in X_N$, $\{y | x \in \text{Supp}(y)\} = \{x\} \cup \{\langle y_1, y_2 \rangle | y_2 = x\}$. Thus we can see that for any $x_1 \in X_S$, the probability of mapping to $\{y | x_1 \in Supp(y)\}$ is always p^*_{LDP}. For any $x_2 \neq x_1$, the probability of mapping to $\{y | x_1 \in Supp(y)\}$ is either q^*_{LDP} or $f(\frac{1}{s}p^*_{LDP} + \frac{s-1}{s}q^*_{LDP})$. As delineated in Eq. (9), the latter probability equates to q^*_{LDP}. For any $x_3 \in X_N$, the probability of mapping to $\{y | x_3 \in Supp(y)\}$ is always $(1-f) + fz$, and no other private data can be mapped to this set. □

The lower bounds of the l_2 error under the LDP model have been proven to be $\Theta(\frac{d}{n\varepsilon^2})$ (when $\varepsilon \in (0,1)$) and $\Theta(\frac{de^\varepsilon}{n(e^\varepsilon-1)^2})$ (when $\varepsilon \in (1, \log d)$) [15], indicating that some LDP mechanisms have achieved optimal utility. A natural question arises: If GLUTF transforms an optimal mechanism under the LDP model, will the resulting ULDP mechanism remain optimal? We found that the order-optimality is also preserved in the GLUTF transformation process.

Theorem 6. *If a pure LDP mechanism is order-optimal, then the ULDP mechanism generated by GLUTF from this LDP mechanism is also order-optimal.*

Proof. From Theorem 4, the MSE of ULDP mechanism generated by GLUTF is $\frac{1}{n}\left((1-\theta)\frac{1-p^*-q^*}{(p^*-q^*)} + s\frac{q^*(1-q^*)}{(p^*-q^*)^2} + \theta\frac{1-z^*}{z^*}\right)$, where the first and second items are the errors caused by sensitive data, and the third item is the error caused by non-sensitive data. We know that $z^* > 1-f = \frac{p^*-q^*}{p^*+q^*(s-1)}$, thus we have $\theta\frac{1-z^*}{z^*} < s\frac{q^*(1-q^*)}{(p^*-q^*)^2}$. This indicates that regardless of the size of sensitive data set s or the frequency of all sensitive data $1-\theta$, the error caused by sensitive data always dominates the MSE. Therefore, if a high data utility LDP mechanism is transformed by GLUTF, the derived ULDP mechanism must have high data utility. The MSE of a pure LDP mechanism [12] can be expressed as: $\frac{1}{n}\left(\frac{1-p^*-q^*}{(p^*-q^*)} + d\frac{q^*(1-q^*)}{(p^*-q^*)^2}\right)$. Specifically, if the MSE of LDP mechanism is $\Theta(\frac{d}{n\varepsilon^2})$ or $\Theta(\frac{de^\varepsilon}{n(e^\varepsilon-1)^2})$, then the MSE of the corresponding ULDP mechanism is $\Theta(\frac{s}{n\varepsilon^2})$ or $\Theta(\frac{se^\varepsilon}{n(e^\varepsilon-1)^2})$, reaching the lower bound of the l_2 error under the ULDP model [9]. This demonstrates that our GLUTF framework preserves order-optimality. □

By substituting p^*, q^*, and z^* into Eq. (28), the MSE of any pure ULDP mechanism can be effectively computed. This not only facilitates the comparative assessment of different pure ULDP mechanisms but also enables the subsequent optimization of specific mechanism parameters. It is noteworthy that the true frequency θ is incorporated in the MSE calculation. This is distinct from some previous works because we use the complete variance, including the term with the true frequency, rather than omitting it. The value of θ can be determined based on some prior knowledge, or estimated in a manner similar to that used by Qian et al [17]. Consequently, θ in Eq. (28) does not impede the design or execution of the mechanism.

Aggregation and utility analysis methods for non-pure mechanisms. When aggregating and analyzing the utility of non-pure ULDP mechanisms, the situation becomes significantly more complex. This complexity arises because the perturbation processes of non-pure ULDP mechanisms differ and cannot be represented by a unified symbol like in pure ULDP. However, we can extract commonalities from the structure of the ULDP mechanism. According to the definition of ULDP, users holding sensitive data are equivalent to executing an LDP mechanism. We denote the aggregation method of this LDP mechanism as *Agg* and the utility analysis method (variance calculation method) as *UA*. In this section, we will continue to use some of the symbols defined in Section 3. However, this will not affect the generality of the conclusions. The following analysis is also applicable to any non-pure ULDP mechanism.

For any non-pure ULDP mechanism, the server can esti-

mate the frequency of private data x as following:

$$\hat{c}_x = \begin{cases} Agg(Y') - \dfrac{f(1-z)}{s}\theta', & \text{if } x \in X_S \\ \dfrac{\sum_{i=1}^{n} \mathbb{1}_x(y_i)}{n(1-f)}, & \text{if } x \in X_N, \end{cases} \tag{30}$$

where $Y' = \{y_i | y_i \in Y_P, 1 \le i \le n\}$, $\theta' = \sum_{x \in X_N} \hat{c}_x$ and $\mathbb{1}_x(y)$ is an indicator function that outputs 1 when $x = y$ and 0 otherwise.

Since the proofs of the following theorems are straightforward, we present the results directly here, with the full proofs provided in Appendix A.

Theorem 7. *If Agg is an unbiased estimation method, then the estimation methods presented in Eq. (30) are also unbiased.*

Theorem 8. *For any non-pure ULDP mechanism, the variance of the estimation $\hat{c}_x$ in Eq. (30) is:*

$$Var[\hat{c}_x] = \begin{cases} UA(Agg(Y')) + \dfrac{f^3(1-z)^2}{ns^2(1-f)}\theta, & \text{if } x \in X_S \\ c_x \dfrac{f}{n(1-f)}, & \text{if } x \in X_N. \end{cases} \tag{31}$$

The inherent flaw of the non-pure ULDP mechanism is its susceptibility to covariance, which makes accurate utility analysis impossible. In Theorem 8, we disregard covariance and obtain an approximate result. Similarly, the uHR [11] protocol also encounters this problem. Although it considers the influence of covariance, it only obtains a loose bound.

5 Mechanism Instantiation

In this section, we will propose three ULDP mechanisms based on some commonly used and highly effective LDP mechanisms (the fourth, uWheel, is detailed in Appendix B as its performance is nearly identical to that of uLH). Subsequently, we will compare the communication cost and data utility of these mechanisms to give the recommended mechanisms for different scenarios. The theoretical analysis results demonstrate that the performance of our proposed mechanisms surpasses that of existing ULDP mechanisms. In GLUTF, we treat the LDP protocol as a black box. Therefore, we will not describe the specific steps of LDP mechanisms in detail in this section.

5.1 uSS

We propose the utility-optimized Subset Selection (uSS) mechanism based on the Subset Selection (SS) mechanism [14, 15]. The SS mechanism is currently one of the best performing mechanisms in the medium privacy regime ($1 < \varepsilon < \log s$) and is a pure LDP mechanism. Given the privacy budget ε and the input domain X_S, in the SS mechanism, it follows that $p^*_{LDP} = \frac{ke^\varepsilon}{ke^\varepsilon+s-k}$, $q^*_{LDP} = \frac{k(ke^\varepsilon+s-k-e^\varepsilon)}{(ke^\varepsilon+s-k)(s-1)}$ and $Y_{LDP} = \{y | y \subseteq X_S, |y| = k\}$. Here, k represents the number of elements in one perturbed data, i.e., $k = |y|$. Subsequently, we will analyze to determine the optimal value of k in the uSS mechanism.

Perturbation. According to Eq. (9), we can calculate that

$$f = \frac{s(e^\varepsilon k - e^\varepsilon - k + s)}{(s-1)(e^\varepsilon k - k + s)}. \tag{32}$$

For any $y_m \in Y_P$, Eq. (15) can be maximized when $x_i \in y_m$. At this point, among p_{tm} $(i = 1, 2, \ldots, s)$, there are k values equal to p_{im} and $s-k$ values equal to $\frac{p_{im}}{e^\varepsilon}$. Thus, we have:

$$\frac{\Pr[\mathcal{A}(x_1) = y]}{\Pr[\mathcal{A}(x_2) = y]} = e^\varepsilon \frac{\frac{p^*_{LDP}}{q^*_{LDP}} + s - 1}{(1-z)(ke^\varepsilon + s - k)} = e^\varepsilon. \tag{33}$$

Therefore, the maximum value of z is $\frac{(e^\varepsilon-1)(k-1)}{e^\varepsilon(k-1)-k+s}$.

Utility analysis. It is evident that in the uSS mechanism, $Y_P = \{y | y \subseteq X_S, |y| = k\}$ and $Y_I = X_N \cup \{\langle y_1, y_2 \rangle | y_1 \in Y_P, y_2 \in X_N\}$. The Supp function for uSS is defined as follows:

$$\text{Supp}(y) = \begin{cases} \{x | x \in y\}, & \text{if } y \in Y_P \cup X_N \\ \{x | x \in y_1\} \cup \{y_2\}, & \text{if } y = \langle y_1, y_2 \rangle. \end{cases} \tag{34}$$

Given the Supp function, the uSS mechanism satisfies ULDP and is pure, with $p^* = \frac{ke^\varepsilon}{ke^\varepsilon+s-k}$, $q^* = \frac{k(ke^\varepsilon+s-k-e^\varepsilon)}{(ke^\varepsilon+s-k)(s-1)}$ and $z^* = \frac{k(e^\varepsilon-1)}{k(e^\varepsilon-1)+s}$. By substituting p^*, q^* and z^* into Eq. (28), we can calculate the MSE of the estimated frequency:

$$\begin{aligned} MSE[\hat{\boldsymbol{c}}] = \frac{1}{n}\Big(s\frac{(ke^\varepsilon - e^\varepsilon + s - k)(ke^\varepsilon - k + s - 1)}{k(s-k)(e^\varepsilon-1)^2} \\ + (1-\theta)\frac{k(1-k)(e^\varepsilon-1) + (s-1)(s-2k)}{k(s-k)(e^\varepsilon-1)} \\ + \theta\frac{s}{k(e^\varepsilon-1)}\Big). \end{aligned} \tag{35}$$

Parameter optimization. We determine the optimal k from a data utility perspective by computing the partial derivative of the MSE:

$$\begin{aligned} \frac{\partial MSE[\hat{\boldsymbol{c}}]}{\partial k} = \frac{1}{k^2 n(e^\varepsilon-1)^2(k-s)^2}\big((e^\varepsilon-1)\theta(k^2(e^\varepsilon(s-1) \\ -1) + 2ks - s^2) - (s-1)^2((e^\varepsilon-1)k + s)(s - (e^\varepsilon+1)k)\big). \end{aligned} \tag{36}$$

By setting $\frac{\partial MSE[\hat{\boldsymbol{c}}]}{\partial k} = 0$, we can obtain the optimal value of k:

$$k = \frac{s}{e^\varepsilon \sqrt{\frac{(s-1)(e^\varepsilon(s+\theta-1)-\theta)}{e^\varepsilon((e^\varepsilon-1)\theta+(s-1)^2)}} + 1}. \tag{37}$$

5.2 uUE

We propose the utility-optimized Unary Encoding (uUE) mechanism based on the Unary Encoding (UE) mechanism [12]. The UE mechanism employs Unary Encoding, which is one of the most common approaches to encoding. Given privacy budget ε and input domain X_S, in the UE mechanism, it follows that $p^*_{LDP} = p$, $q^*_{LDP} = \frac{p}{e^\varepsilon(1-p)+p}$ and $Y_{LDP} = \{0,1\}^s$, a binary bit vector of size s. Subsequently, we will analyze to determine the optimal value of p in the uUE mechanism.

Perturbation. According to Eq. (9), we can calculate that

$$f = \frac{s}{s+(e^\varepsilon-1)(1-p)}. \tag{38}$$

For any $y_m \in Y_P$, we denote the number of 1 in the vector y as j $(0 \le j \le s)$. It is evident that Eq. (15) reaches its global maximum when $j = 1$ and the corresponding position of x_i in y_m is 1, so we have:

$$\frac{\Pr[\mathcal{A}(x_1)=y]}{\Pr[\mathcal{A}(x_2)=y]} = e^\varepsilon \frac{\frac{p^*_{LDP}}{q^*_{LDP}}+s-1}{(1-z)(e^\varepsilon+s-1)} = e^\varepsilon. \tag{39}$$

Therefore, the maximum value of z is $\frac{p(e^\varepsilon-1)}{e^\varepsilon+s-1}$.

Utility analysis. It is evident that in the uUE mechanism, $Y_P = \{0,1\}^s$ and $Y_I = X_N \cup \{\langle y_1,y_2\rangle | y_1 \in Y_P, y_2 \in X_N\}$. The Supp function for uUE is defined as follows:

$$\text{Supp}(y) = \begin{cases} \{x|y[x]=1\}, & \text{if } y \in Y_P \\ \{x\}, & \text{if } y \in X_N \\ \{x|y_1[x]=1\} \cup \{y_2\}, & \text{if } y = \langle y_1,y_2\rangle. \end{cases} \tag{40}$$

Given the Supp function, the uUE mechanism satisfies ULDP and is pure, with $p^* = p$, $q^* = \frac{p}{e^\varepsilon(1-p)+p}$ and $z^* = \frac{e^\varepsilon-1}{e^\varepsilon+s-1}$. By substituting p^*, q^*, and z^* into Eq. (28), we can calculate the MSE of the estimated frequency:

$$MSE[\hat{\boldsymbol{c}}] = \frac{1}{n}(s\frac{e^\varepsilon}{p(1-p)(e^\varepsilon-1)^2} + (1-\theta)\frac{p^2-e^\varepsilon(1-p)^2}{p(p-1)(e^\varepsilon-1)} + \theta\frac{s}{e^\varepsilon-1}) \tag{41}$$

Parameter optimization. We determine the optimal p from a data utility perspective by computing the partial derivative of the MSE:

$$\frac{\partial MSE[\hat{\boldsymbol{c}}]}{\partial p^*} = \frac{1}{n(e^\varepsilon-1)^2(p^*-1)^2(p^*)^2}(e^{2\varepsilon}(p^*-1)^2(\theta-1) + e^\varepsilon(2p^*-1)(s+\theta-1) - (p^*)^2(\theta-1)). \tag{42}$$

By setting $\frac{\partial MSE[\hat{\boldsymbol{c}}]}{\partial k} = 0$, we can obtain the optimal value of p:

$$p = \frac{1}{\sqrt{\frac{e^\varepsilon s+(e^\varepsilon-1)(\theta-1)}{e^\varepsilon(s-(e^\varepsilon-1)(\theta-1))}}+1}. \tag{43}$$

5.3 uLH

We propose the utility-optimized Local Hashing (uLH) mechanism based on the Local Hashing (LH) mechanism [12]. The LH mechanism employs the Hash function to encode private data, thereby achieving high data utility and low communication cost, especially when dealing with large-dimensional private data. Given privacy budget ε and input domain X_S, in the LH mechanism, it follows that $p^*_{LDP} = \frac{e^\varepsilon}{e^\varepsilon+g-1}$, $q^*_{LDP} = \frac{1}{g}$ and $Y_{LDP} = \{\langle H,y\rangle | H \in \mathcal{H}, y \in \{1,2,\ldots,g\}\}$. Here, $\mathcal{H}$ denotes a collection of hash functions, with each element having a value domain $\{1,2,\ldots,g\}$. Subsequently, we will analyze what the optimal value for g is.

Perturbation. According to Eq. (9), we can calculate that

$$f = \frac{s(e^\varepsilon+g-1)}{e^\varepsilon g+(e^\varepsilon+g-1)(s-1)}. \tag{44}$$

For any $\langle H,y_m\rangle \in Y_P$, Eq. (15) can be maximized when $H(x_i) = y_m$, in which case $p_{im} = \frac{e^\varepsilon}{e^\varepsilon+g-1}$ and $p_{tm} = \frac{1}{g}$ $(1 \le t \le s, t \ne i)$. Thus, we have:

$$\frac{\Pr[\mathcal{A}(x_1)=y]}{\Pr[\mathcal{A}(x_2)=y]} = e^\varepsilon \frac{g}{(1-z)(e^\varepsilon+g-1)} = e^\varepsilon. \tag{45}$$

Therefore, the maximum value of z is $\frac{e^\varepsilon-1}{e^\varepsilon+g-1}$.

Utility analysis. It is evident that in the uLH mechanism, $Y_P = \{\langle H,y\rangle | H \in \mathcal{H}, y \in \{1,2,\ldots,g\}\}$ and $Y_I = X_N \cup \{\langle y_1,y_2\rangle | y_1 \in Y_P, y_2 \in X_N\}$. The Supp function for uLH is defined as follows:

$$\text{Supp}(y) = \begin{cases} \{x|H(x)=y'\}, & \text{if } y = \langle H,y'\rangle \in Y_P \\ \{y\} & \text{if } y \in X_N \\ \{x|H(x)=y_1\} \cup \{y_2\}, & \text{if } y = \langle\langle H,y_1\rangle,y_2\rangle. \end{cases} \tag{46}$$

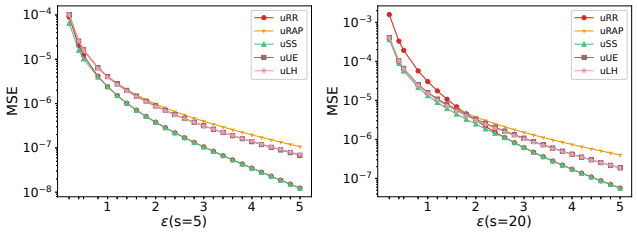
Given the Supp function, the uLH mechanism satisfies ULDP and is pure, with $p^* = \frac{e^\varepsilon}{e^\varepsilon+g-1}$, $q^* = \frac{1}{g}$ and $z^* = \frac{(e^\varepsilon-1)(g+s-1)}{e^\varepsilon(g+s-1)+(g-1)(s-1)}$. By substituting p^*, q^*, and z^* into Eq. (28), we can calculate the MSE of the estimated frequency:

$$MSE[\hat{\boldsymbol{c}}] = \frac{1}{n}(s\frac{(e^\varepsilon+g-1)^2}{(e^\varepsilon-1)^2(g-1)} + (1-\theta)\frac{(g-1)^2-e^\varepsilon}{(e^\varepsilon-1)(g-1)} + \theta\frac{gs}{(e^\varepsilon-1)(g+s-1)}). \tag{47}$$

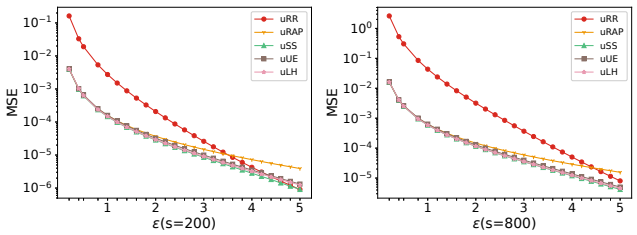
Parameter optimization. Considering the application scenario of the uLH mechanism, it is known that $s \gg g-1$. To simplify the process of calculating partial derivatives and to avoid overly complex results, we let $MSE \approx \frac{1}{nd}(s\frac{(e^\varepsilon+g-1)^2}{(e^\varepsilon-1)^2(g-1)} + (1-\theta)\frac{(g-1)^2-e^\varepsilon}{(e^\varepsilon-1)(g-1)} + \theta\frac{g}{(e^\varepsilon-1)})$. We determine the optimal g from a data utility perspective by comput-

Table 1: Comparison of communication costs and MSE among various mechanisms.

Mechanism	Communication cost	High privacy regime ($0 < \varepsilon < 1$)		Medium privacy regime ($1 < \varepsilon < \log s$)	
		MSE	Order-optimal	MSE	Order-optimal
uRR	$O(\log d)$	$O(\frac{s^2}{n\varepsilon^2})$	No	$O(\frac{s^2}{n(e^\varepsilon-1)^2})$	No
uRAP	$O(\log(2^s+d-s))$	$O(\frac{s}{n\varepsilon^2})$	Yes	$O(\frac{se^{\varepsilon/2}}{n(e^{\varepsilon/2}-1)^2})$	No
uSS	$O(\log(C_s^{\frac{s}{e^\varepsilon+1}}+d-s))$	$O(\frac{s}{n\varepsilon^2})$	Yes	$O(\frac{se^\varepsilon}{n(e^\varepsilon-1)^2})$	Yes
uUE	$O(\log(2^s+d-s))$	$O(\frac{s}{n\varepsilon^2})$	Yes	$O(\frac{se^\varepsilon}{n(e^\varepsilon-1)^2})$	Yes
uLH	$O(\log(e^\varepsilon+d-s))$	$O(\frac{s}{n\varepsilon^2})$	Yes	$O(\frac{se^\varepsilon}{n(e^\varepsilon-1)^2})$	Yes



(a) $s = 5$ (b) $s = 20$



(c) $s = 200$ (d) $s = 800$

Figure 3: Theoretical evaluation results of MSE for each ULDP mechanism in different scenarios.

ing the partial derivative of the MSE:

$$\frac{\partial MSE[\hat{c}]}{\partial g} = \frac{1}{n(e^\varepsilon-1)^2(g-1)^2}(e^\varepsilon((g-2)g+\theta) + (g-1)^2(s-1) - (e^{2\varepsilon}(s+\theta-1))). \quad (48)$$

By setting $\frac{\partial MSE[\hat{c}]}{\partial k} = 0$, we can obtain the optimal value of g:

$$g = e^\varepsilon \sqrt{\frac{e^\varepsilon(s+\theta-1)-\theta+1}{e^\varepsilon(e^\varepsilon+s-1)}} + 1. \quad (49)$$

5.4 Comparison of the mechanisms

We have converted several commonly used LDP mechanisms in the domain of frequency estimation to ULDP mechanisms. The communication costs and MSE of these mechanisms have been analyzed, with the findings presented in Tab. 1.

Communication costs. Based on Definition 5, the communication cost of a mechanism can be expressed as $O(\log|Y|)$. For uRR and uRAP, $|Y|$ is d and (2^s+d-s), respectively. For other mechanisms, $|Y|$ is $(|Y_P|+|X_N|+|Y_P|\cdot|X_N|)$. We adopt the relaxed bound $O(\log(|Y_P|+|X_N|))$ for simplicity, since $\log(|Y_P|+|X_N|+|Y_P|*|X_N|) < 2*\log(|Y_P|+|X_N|)$.

Tab. 1 presents the existing ULDP mechanisms: uRR and uRAP (uHR is excluded from comparison due to its loose bound on data utility), along with our three proposed ULDP mechanisms: uSS, uUE, and uLH. It is evident that our proposed ULDP mechanisms achieve order-optimality in both high and medium privacy regimes, a feat that existing ULDP mechanisms cannot accomplish. When s is large, our proposed uLH mechanism also achieves the lowest communication cost (with the optimal value of g approximating $(e^\varepsilon+1)$).

We have selected a range of scenarios, showcasing the theoretical MSE of each mechanism, as depicted in Fig. 3. It is evident that the uSS mechanism serves as an optimized variant of the uRR mechanism. Similarly, the uUE mechanism can be regarded as an optimized version of the uRAP mechanism. In light of these observations and the data presented in Fig. 3, we propose the following recommendations:

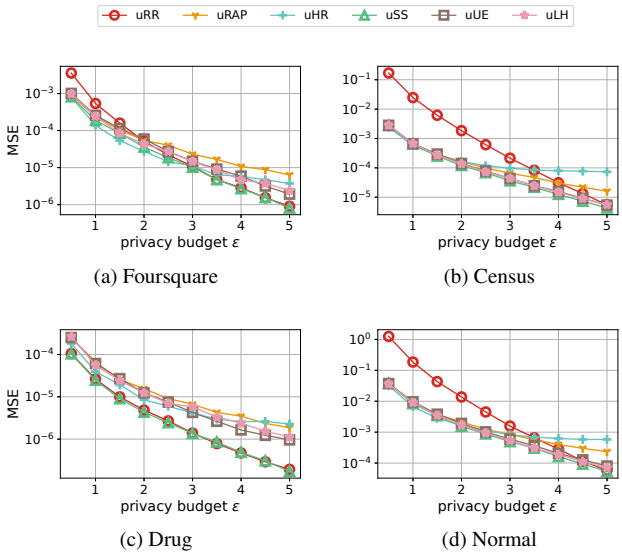
- **Low-sensitive data scenarios.** When s is small, the uSS mechanism demonstrates optimal data utility coupled with low communication cost, making it our recommended choice.

- **High-sensitive data scenarios.** When s is large, the uSS, uUE and uLH mechanisms exhibit very similar data utility. Considering the communication cost, our recommendation leans towards uLH mechanism.

- **Communication-cost prioritized scenarios.** The uLH protocol incorporates a hash function in its perturbation process, significantly reducing communication costs when s is large, making it our top recommendation. However, when s is small, the hash function provides no significant benefit, and we instead recommend the uSS and uUE mechanisms.

- **Computation-cost prioritized scenarios.** The aggregation process of uLH requires executing a large number of hash functions, which significantly increases computational overhead. In such cases, we recommend the uSS and uUE mechanisms.

Collaborative sampling. The parameter θ, integral to both the MSE and mechanism parameters, represents the sum of the true frequencies of all non-sensitive data. In practical scenarios, neither users nor servers can ascertain the exact value of θ. While a priori knowledge might assist in estimating θ, the mechanism is designed to function effectively even in its absence. A feasible solution to this challenge is collaborative sampling [17]. Initially, θ is assigned an arbitrary value, followed by sampling approximately 5% of users to approximate θ. This estimated value of θ is then employed in the regular frequency estimation process for the remaining users. The efficacy of this approach is validated by the experimental results in Section 6.3.



(a) Foursquare (b) Census

(c) Drug (d) Normal

Figure 4: The impact of privacy budget ε on the performance of each ULDP mechanism under different datasets.

6 Experimental Evaluation

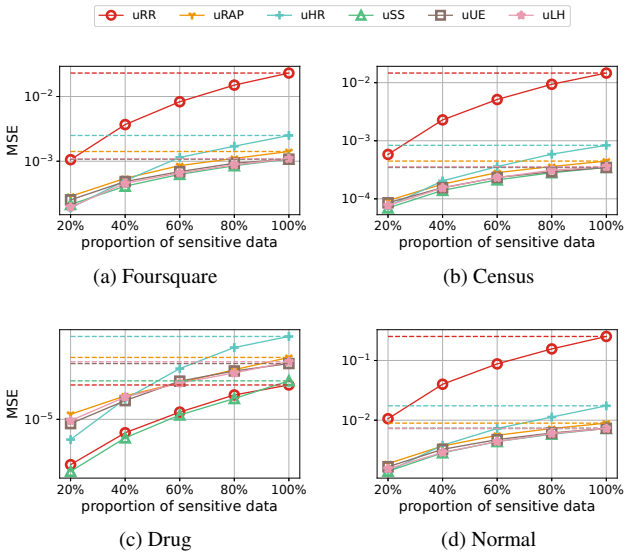
6.1 Experimental Set-up

We conducted experiments on both real and synthetic datasets.
Foursquare dataset. The Foursquare dataset (Global-scale check-in dataset) [18, 19] contains 33,278,683 global check-in records, each associated with a POI ID and venue type. We used 419,959 records from the Manhattan area for our experiments. Manhattan was divided into a 25×25 grid ($|X| = 625$), and regions containing hospitals, casinos, or strip clubs visited by at least 10 users were considered sensitive ($s = 25$).
Census dataset. The Census dataset [20] contains 2,458,285 records from the U.S. census, with each record containing 68 attributes. We used all users in the dataset and selected age, income, marital status, sex, and disability as attributes, with 8, 5, 5, 2, and 3 categories, respectively ($|X| = 8*5*5*2*3 = 1200$). Categories involving divorce, unemployment, or disability were considered sensitive. As users in the child age group are not associated with these sensitive attributes (details in [20]), some categories were excluded, resulting in $s = 424$.
Drug dataset. The Drug dataset [21] contains 215,063 patient reviews of specific drugs, each with 6 attributes. We used all users in the dataset and selected the rating attribute, which has 10 categories ($|X| = 10$). Low ratings were considered sensitive ($s = 3$).
Normal dataset. The Normal dataset contains 99,732 simulated records following a normal distribution, with a total domain of 1000 ($|X| = 1000$). We applied the average sensitive proportion from other datasets (23%) to randomly select sensitive data, resulting in $s = 230$.
Evaluation details. To mitigate the impact of randomness on our experimental results, each experiment is repeated 50 times, and the average is taken as the final result. The MSE of the mechanism is calculated using the formula $\frac{1}{m}\sum_{i=1}^{m}(c_i(x) - \hat{c}_i(x))^2$, where m represents the number of repetitions. During the execution of the uLH mechanism, $|\mathcal{H}| = 100000$.
Mechanisms for comparison. The experiment involves six mechanisms in total, among which uRR [9], uRAP [9], and uHR [11] are existing ULDP mechanisms, while uSS, uUE, and uLH are the mechanisms proposed in Section 5.

6.2 Impact of Parameters

The impact of the privacy budget. We conducted experiments to assess the impact of ε on mechanism performance. The experimental results are shown in Fig. 4. It is observed that the MSE of each mechanism decreases with an increase in ε. Additionally, comparing Fig. 4(a), Fig. 4(b), Fig. 4(c) and Fig. 4(d) reveals a significant shift in the performance ranking of the mechanisms across different datasets (essentially, different s values).

The uSS consistently exhibits the highest data utility across various datasets. In contrast, the performance of the uRR differs significantly across datasets. Specifically, under the Drug dataset, the uRR ranks among those with the highest data utility, whereas it shows the lowest utility in the Census and Normal dataset. This difference can be further explained by the change in the optimal k (a parameter in the uSS) across scenarios. Notably, the uRR is a special case of the uSS when $k = 1$. As s increases, the optimal k diverges from 1, resulting in degraded performance of the uRR. As ε increases, the optimal k approaches 1, resulting in improved performance ranking of the uRR mechanism.



(a) Foursquare (b) Census

(c) Drug (d) Normal

Figure 5: The impact of sensitive data proportion on the performance of each ULDP mechanism under different datasets.

The uUE and the uRAP use similar encoding methods, but the former consistently achieves higher data utility than the latter. This can be attributed to shifts in the optimal p. For example, in Census dataset, when $\varepsilon = 0.5$, the p for uRAP and uUE are 0.562 and 0.5, respectively; and when $\varepsilon = 5$, the p are 0.924 and 0.517, respectively. This demonstrates that as ε increases, the p used by the uRAP increasingly deviates from its optimal value.

The data utility of uLH consistently resembles that of uUE, as expected because "OLH can be viewed as a compact way of implementing OUE" [12]. Therefore, it is reasonable that their corresponding ULDP versions exhibit similar performance.

A surprising observation is that as ε increases, the MSE variation for uHR diminishes, even remaining constant when $\varepsilon > 3.5$. We cannot theoretically predict this phenomenon because uHR cannot obtain an accurate theoretical MSE due to its unique perturbation mechanism. In practice, this poses a significant drawback, as it hinders the selection of a suitable privacy budget and prevents reliable error estimation.

The impact of sensitive data proportion and ULDP vs. LDP. We conducted experiments to assess the impact of sensitive data proportion on mechanism performance. We set $\varepsilon = 2$. The experimental results are shown in Fig. 5. The results show that performance declines with increasing sensitive data proportion, consistent with theoretical predictions. It can be observed that as the sensitive data proportion increases, the MSE of the uHR increases more rapidly than other mechanisms. This suggests that the uHR mechanism is not recommended when the proportion of sensitive data is high ($\frac{s}{d} \geq 40\%$).

Notably, when the sensitive data proportion reaches 100%, the six ULDP mechanisms essentially operate as LDP mechanisms, yielding the lowest utility. This supports the idea that transforming LDP into ULDP addresses sensitivity distinction and improves data utility.

Table 2: Error assessment between the true frequency θ and the estimated frequency $\hat{\theta}$.

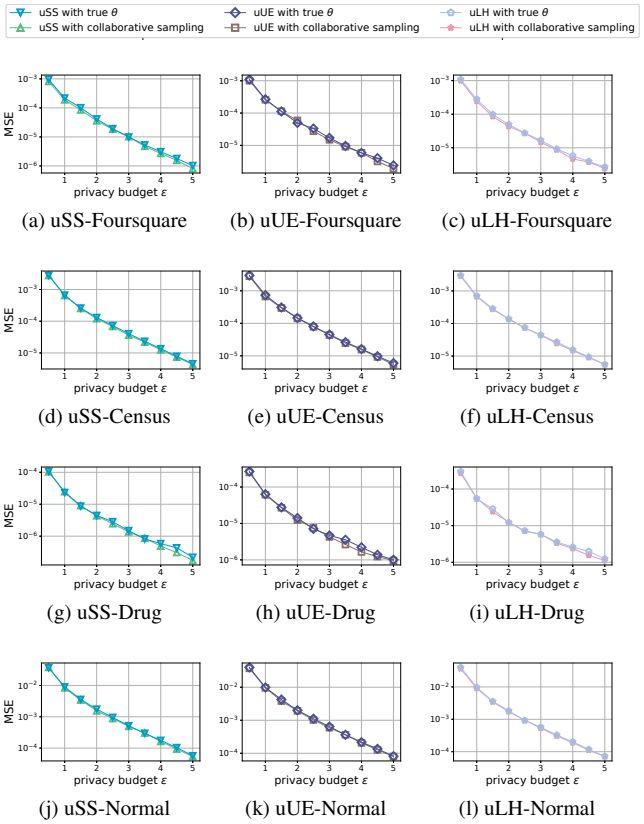
	Foursquare	Census	Drug	Normal
uSS	1.23%	0.40%	1.51%	2.45%
uUE	3.35%	4.38%	1.79%	18.09%
uLH	0.83%	0.41%	1.26%	2.90%

Experimental settings highlighting the utility advantage of our mechanisms. We observe that uSS consistently outperforms uRR, and uUE consistently outperforms uRAP across all settings. This advantage becomes more pronounced as $|\frac{s}{e^{\varepsilon}+1} - 1|$ and $|0.5 - \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}|$ increase. While uHR performs similarly to our mechanisms at small privacy budgets, its performance degrades significantly as the budget increases. In terms of data utility, uSS consistently achieves the best performance. In terms of communication cost, uLH offers the lowest cost while maintaining high data utility when s is large.

6.3 Effect of Collaborative Sampling

We evaluated the efficacy of cooperative sampling without prior knowledge. For the Foursquare, Census, Drug, and Normal datasets, the actual θ values are 0.95, 0.56, 0.78, and 0.77, respectively. In our experimental setup, we initially set $\theta = 0$ and randomly selected 5% of users to implement the mechanisms, yielding an estimate $\hat{\theta}$. The privacy budget ε was varied from 0.5 to 5. The accuracy of $\hat{\theta}$ is measured by $\frac{|\theta - \hat{\theta}|}{\theta}$, with results shown in Tab. 2. The findings reveal that the discrepancy between $\hat{\theta}$ and the true θ is less than 4.38% for all mechanisms except uUE, which shows significant deviation on the Normal dataset. This inaccuracy aligns with theoretical expectations. It is evident that a larger z^* leads to a more accurate estimate of θ. For example, when $\varepsilon = 0.5$ in the Normal dataset, the z^* values for uSS, uUE, and uLH are 0.20, 0.0028 and 0.18, respectively.

Then, $\hat{\theta}$ was used to estimate the frequency of the remaining 95% of users. In the same setting, we executed the mechanisms using the true θ and 100% of users, and compared the results, as shown in Fig. 6. The findings indicate that the MSE of cooperative sampling is close to the MSE using the true θ, across various datasets and ε. This demonstrates that the mechanism remains effective even without prior knowledge. This effectiveness can be attributed to two factors: a relatively accurate θ can still be achieved with just 5% of users, and θ precision has minimal impact on results, as shown by uUE performance on the Normal dataset.



(a) uSS-Foursquare (b) uUE-Foursquare (c) uLH-Foursquare

(d) uSS-Census (e) uUE-Census (f) uLH-Census

(g) uSS-Drug (h) uUE-Drug (i) uLH-Drug

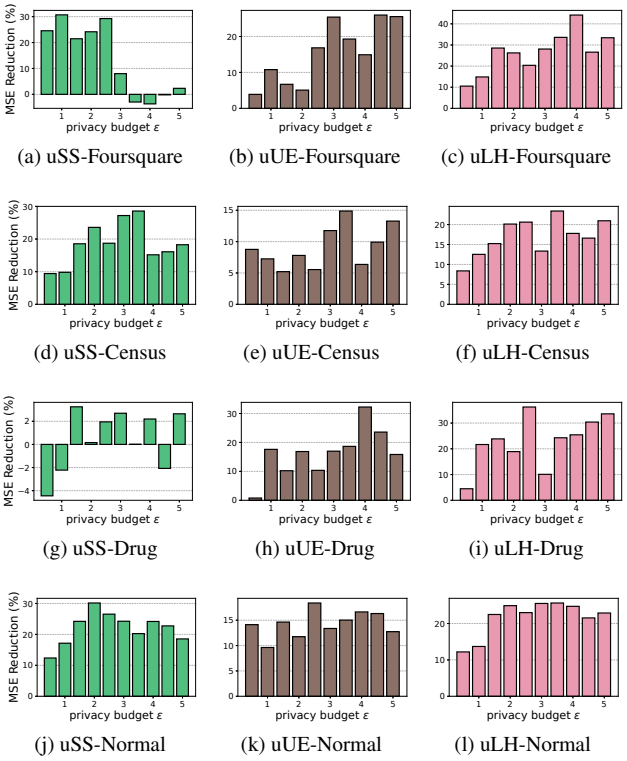
(j) uSS-Normal (k) uUE-Normal (l) uLH-Normal

Figure 6: The impact of privacy budget ε on the performance of each ULDP mechanism under different datasets by using estimated frequency $\hat{\theta}$ or true frequency θ.

6.4 Evaluation of Maximizing z

In Section 3, we advocated the maximization of the value z to enhance data utility. We empirically examine whether this practice effectively reduces the MSE. We compared the cases where z was set as normal and where it was forced to zero, with the corresponding MSEs denoted as mse_1 and mse_2, respectively. We used metric $\frac{mse_2 - mse_1}{mse_2}$ to determine whether the MSE was reduced. The experimental results are presented in Fig. 7. We observed that across different datasets, privacy budgets, and mechanisms, the MSE decreased by an average of $9\% \sim 26\%$. This indicates that maximizing z in GLUTF can indeed significantly reduce the MSE.

The results in Fig. 7(a) and 7(g) differ significantly from those of other settings. At certain privacy budgets, the reduction in MSE is minimal and may even increase. This is consistent with theoretical expectations. For the uSS, when applied to the Foursquare dataset ($\varepsilon > 3.5$) or the Drug dataset ($\varepsilon > 0$), the optimal k is 1, resulting in the maximum possible z being 0. It should be noted that the increase in MSE is not caused by our mechanism, but rather by the noise introduced by random perturbation.



(a) uSS-Foursquare (b) uUE-Foursquare (c) uLH-Foursquare

(d) uSS-Census (e) uUE-Census (f) uLH-Census

(g) uSS-Drug (h) uUE-Drug (i) uLH-Drug

(j) uSS-Normal (k) uUE-Normal (l) uLH-Normal

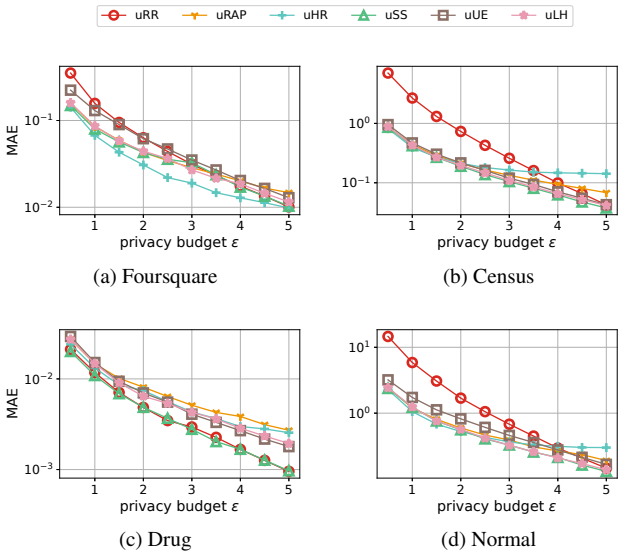
Figure 7: The impact of maximizing z on the performance of each ULDP mechanism under different datasets.

6.5 Extended utility metrics

In the preceding experiments, we used MSE to evaluate the mechanisms' performance. While MSE is a reasonable metric, it is not sufficient. Since our mechanisms are theoretically optimized for MSE, their superior performance on this metric is unsurprising. To multi-perspectively assess the performance of our mechanisms, we extend utility metrics.

Evaluation on MAE. We used Mean Absolute Error (MAE, computed as $\frac{1}{m}\sum_{i=1}^{m}|c_i(x) - \hat{c}_i(x)|$) to evaluate the utility of the mechanisms, with the results shown in Fig. 8. Many conclusions drawn in Section 6.2 still hold: uSS remains superior to uRR, and uUE and uLH exhibit similar utility. However, two surprising observations were made. First, in Fig. 8(a) and 8(d), uUE does not consistently outperform uRAP. This is because the parameter p in uUE was optimized to minimize MSE, which, to some extent, caused an "overfitting" to MSE. Second, in Fig. 8(a), uHR emerged as the best-performing mechanism under MAE, in contrast to its performance under MSE. This difference occurs because large errors, which are more common in uHR, are more heavily penalized in MSE.

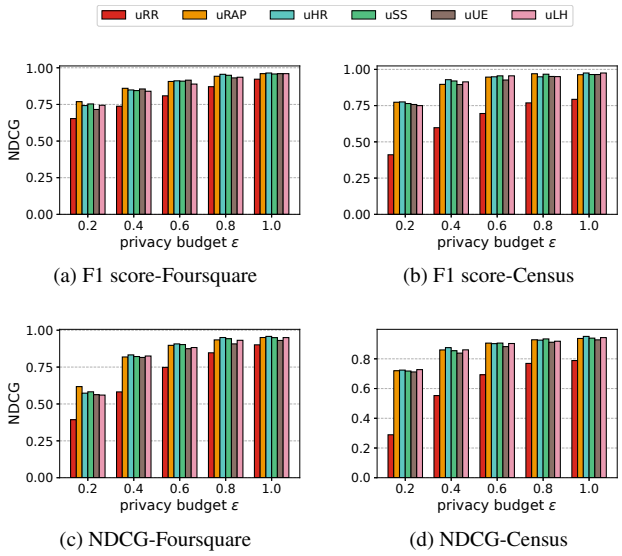
Evaluation on frequent item mining. We conducted frequent item mining experiments to identify the top 30 most frequent data items, using F1 [7] and NDCG [22] as metrics. F1 measures the accuracy of identified items, while NDCG



(a) Foursquare (b) Census

(c) Drug (d) Normal

Figure 8: Performance of ULDP mechanisms on MAE.



(a) F1 score-Foursquare (b) F1 score-Census

(c) NDCG-Foursquare (d) NDCG-Census

Figure 9: Performance of ULDP mechanisms on frequent item mining.

focuses on ranking. As shown in Fig. 9, uRR performs the worst because its estimation error increases with s, distorting the frequency ranking. In contrast, other mechanisms perform similarly, as F1 and NDCG are less sensitive to absolute frequency errors and mainly reflect relative ranking, which stays relatively stable across mechanisms. This suggests placing more emphasis on non-utility factors, such as communication and computation cost, in real-world ULDP frequent item mining.

7 Related work

Frequency estimation for categorical data is one of the fundamental tasks under LDP [3]. Various mechanisms have been proposed for LDP frequency estimation, including GRR [13], RAPPOR [4], SS [14, 15], OUE [12], OLH [12] and Wheel [16]. However, LDP assumes all private data are equally sensitive, leading to excessive obfuscation and potential loss of utility.

To address the issue of sensitivity distinction, several variants of LDP have been proposed. The ULDP [9] model takes into account the difference in data sensitivity and greatly improves the data utility. However, its full potential has yet to be reached (see Section 1). [11] proposed the High-Low LDP model, which is fundamentally equivalent to ULDP. Some researchers have proposed assigning different sensitivity levels to each data. Motivated by this idea, [10] proposed ID-LDP model, and [23] proposed IPLDP. However, assigning different sensitivity levels to each data item is a significant challenge in real-world scenarios, as it is difficult for all users to reach a consensus on the sensitivity ranking of the data. Furthermore, ID-LDP offers lower data utility compared to ULDP [24], while IPLDP can only be applied to the same direct encoding method as in GRR. We believe that among the aforementioned models, only ULDP achieves the most balanced performance, which also motivates our work.

Under the ULDP model, mechanisms such as uRR [9], uRAP [9] and uHR [11] have been proposed for frequency estimation of categorical data, which is the focus of this study. However, they either perform poorly when the sensitive domain is large or fail to obtain accurate theoretical error bounds (see Section 1). [25, 26] extended the ULDP model and proposed (ε, δ)-ULDP model. However, this model does not provide strict LDP-level privacy protection for sensitive data. [27] extends the ULDP model to key-value data and proposes the UKVLDP model. This work is orthogonal to ours.

8 Conclusion

To systematically address the issue of sensitivity distinction in the LDP model, we propose the GLUTF framework that enables the conversion of any LDP mechanism into its corresponding ULDP mechanism while preserving order-optimality and unbiased estimation. We propose the pure ULDP framework and develop a general aggregation and utility analysis method applicable to all ULDP mechanisms generated by GLUTF. We also propose three new ULDP mechanisms and demonstrate through theoretical analysis and experimental validation that these mechanisms achieve better utility than existing ULDP mechanisms. This work establishes a substantial bridge between LDP model and ULDP model.

Acknowledgments

We thank the anonymous shepherd and reviewers for their insightful suggestions and comments. This work is partially supported by the National Natural Science Foundation of China (No. 62172216, No. 62261160651, No. 62402147), the National Key R&D Program of China (No. 2021YFB3100400), and the Fundamental Research Funds for the Central Universities (No. NP2024117).

Ethical Considerations

General Ethical Compliance. Our study adheres to ethical guidelines and best practices in data privacy and security. We ensure that all methodologies comply with established ethical standards to prevent potential risks related to data misuse.
Privacy and Data Protection. Since our research involves differential privacy techniques, we emphasize privacy preservation in data collection, processing, and analysis. The mechanisms we propose are designed to minimize the risk of individual data exposure while maintaining statistical utility.
Human Subjects and Informed Consent. This study does not involve direct human participation. Any datasets used in our experiments are either publicly available or synthetically generated. If real-world data is utilized, it is fully anonymized, and no personally identifiable information is included.
Bias and Fairness. We acknowledge the potential biases that may arise when implementing privacy-preserving mechanisms. We strive to mitigate these biases by thoroughly evaluating the impact of different privacy parameters on data utility across various scenarios.

Compliance with the open science policy

To promote transparency and reproducibility, we commit to making our implementation code, experimental scripts, and datasets publicly available. Accordingly, they have been deposited at https://zenodo.org/records/15614307 under an open-source license.

References

[1] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *The Third Theory of Cryptography Conference, New York, NY, USA, March 4-7*, pages 265–284. Springer, 2006.

[3] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE, 2013.

[4] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[5] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–10. Curran Associates, Inc., 2017.

[6] ADP Team et al. Learning with privacy at scale. *Apple Mach. Learn. J*, 1(8):1–25, 2017.

[7] Tianhao Wang, Ninghui Li, and Somesh Jha. Locally differentially private heavy hitter identification. *IEEE Transactions on Dependable and Secure Computing*, 18(2):982–993, 2019.

[8] Kai Dong, Zheng Zhang, Chuang Jia, Zhen Ling, Ming Yang, Junzhou Luo, and Xinwen Fu. Relation mining under local differential privacy. In *33rd USENIX Security Symposium*, pages 955–972, 2024.

[9] Takao Murakami and Yusuke Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation. In *28th USENIX Security Symposium*, pages 1877–1894, 2019.

[10] Xiaolan Gu, Ming Li, Li Xiong, and Yang Cao. Providing input-discriminative protection for local differential privacy. In *IEEE 36th International Conference on Data Engineering*, pages 505–516. IEEE, 2020.

[11] Jayadev Acharya, Kallista Bonawitz, Peter Kairouz, Daniel Ramage, and Ziteng Sun. Context aware local differential privacy. In *International Conference on Machine Learning*, pages 52–62. PMLR, 2020.

[12] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium*, pages 729–745, 2017.

[13] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[14] Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*, 2016.

[15] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.

[16] Shaowei Wang, Yuqiu Qian, Jiachun Du, Wei Yang, Liusheng Huang, and Hongli Xu. Set-valued data publication with local privacy: tight error bounds and efficient mechanisms. *Proceedings of the VLDB Endowment*, 13(8):1234–1247, 2020.

[17] Qiuyu Qian, Qingqing Ye, Haibo Hu, Kai Huang, Tom Tak-Lam Chan, and Jin Li. Collaborative sampling for partial multi-dimensional value collection under local differential privacy. *IEEE Transactions on Information Forensics and Security*, 18:3948–3961, 2023.

[18] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns. *Journal of Network and Computer Applications*, 55:170–180, 2015.

[19] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology*, 7(3):1–23, 2016.

[20] Chris Meek, Bo Thiesson, and David Heckerman. US Census Data (1990). UCI Machine Learning Repository, 2001. DOI: https://doi.org/10.24432/C5VP42.

[21] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 international conference on digital health*, pages 121–125, 2018.

[22] Xiaochen Li, Weiran Liu, Jian Lou, Yuan Hong, Lei Zhang, Zhan Qin, and Kui Ren. Local differentially private heavy hitter detection in data streams with bounded memory. *Proceedings of the ACM on Management of Data*, 2(1):1–27, 2024.

[23] Xin Li, Hong Zhu, Zhiqiang Zhang, and Meiyi Xie. Item-oriented personalized ldp for discrete distribution estimation. In *European Symposium on Research in Computer Security*, pages 446–466. Springer, 2023.

[24] Takao Murakami and Yuichi Sei. Automatic tuning of privacy budgets in input-discriminative local differential privacy. *IEEE Internet of Things Journal*, 10(18):15990–16005, 2023.

[25] Yue Zhang, Youwen Zhu, Yuqian Zhou, and Jiabin Yuan. Frequency estimation mechanisms under (ε,δ)-utility-optimized local differential privacy. *IEEE Transactions on Emerging Topics in Computing*, 12(1):316–327, 2023.

[26] Yue Zhang, Youwen Zhu, Shaowei Wang, and Xiaohua Huang. Mean estimation of numerical data under (ε,δ)-utility-optimized local differential privacy. *IEEE Transactions on Information Forensics and Security*, 19:9656–9669, 2024.

[27] Bin Wang, Chao Yang, and Jianfeng Ma. UKVLDP: Utility-optimized local differential privacy mechanism for key-value iot data collection. *IEEE Internet of Things Journal*, 12(11):17964–17976, 2025.

Appendices

A Proof for Theorem 7 and Theorem 8

The following is the proof of Theorem 7.

Proof. For $x \in X_N$, we have:

$$E[\hat{c}_x] = \frac{nc_x(1-f)}{n(1-f)} = c_x. \tag{50}$$

For $x \in X_S$, we have:

$$E[\hat{c}_x] = c_x + \frac{f(1-z)}{s}\theta - \frac{f(1-z)}{s}\theta = c_x. \tag{51}$$

In summary, $\hat{c}_x$ is an unbiased estimate of c_x. □

The following is the proof of Theorem 8.

Proof. For non-sensitive data, the variance can be calculated as follows:

$$Var[\hat{c}_x] = Var[\frac{\sum_{i=1}^{n}\mathbb{1}_x(y_i)}{n(1-f)}] = c_x\frac{f}{n(1-f)} \tag{52}$$

In Eq. (30), we use θ' to approximate the proportion of users holding non-sensitive data. First, let's analyze its variance:

$$Var[\theta'] = Var[\sum_{x\in X_N}\hat{c}_x] = \theta\frac{f}{n(1-f)} \tag{53}$$

Now, we can calculate the variance of the estimated frequency of the sensitive data.

$$Var[\hat{c}_x] \approx UA(Agg(Y')) + \frac{f^2(1-z)^2}{s^2}Var[\theta'] \tag{54}$$

$$= UA(Agg(Y')) + \frac{f^3(1-z)^2}{ns^2(1-f)}\theta \tag{55}$$

□

B uWheel

We propose the uWheel mechanism based on the Wheel mechanism. The Wheel mechanism also employs a hash function to convert private data into numerical data within $[0,1)$, and has similar performance to the OLH mechanism. Given privacy budget ε and input domain X_S, in the Wheel mechanism, it follows that $p^*_{LDP} = \frac{le^\varepsilon}{le^\varepsilon - l + 1}$, $q^*_{LDP} = l$ and $Y_{LDP} = \{\langle H, y\rangle | H \in \mathcal{H}, y \in [0,1)\}$. Here, $\mathcal{H}$ denotes a collection of hash functions, where each $H \in \mathcal{H}$ outputs a value in the range of $[0,1)$. Additionally, l is defined as a coverage parameter, which is used to adjust the true/false coverage probability of an item.

Encoding. According to Eq. (9), we can calculate that

$$f = \frac{s(le^\varepsilon + 1 - l)}{e^\varepsilon + (s-1)(le^\varepsilon + 1 - l)}. \tag{56}$$

Then, for any $x \in X_N$, It will be transformed into sensitive data or keep itself.

Next, for $x \in X_S$, the user randomly selects a hash function $H \in \mathcal{H}$ to hash x into $[0,1)$. Conversely, for $x \in X_N$, it remains unchanged.

$$\dot{x} = Encode(x) = \begin{cases} \langle H, H(x)\rangle, & \text{if } x \in X_S \\ x, & \text{if } x \in X_N. \end{cases} \tag{57}$$

Perturbation. It is evident that in the uWheel mechanism, $Y_P = \{\langle H, y\rangle | H \in \mathcal{H}, y \in [0,1)\}$ and $Y_I = X_N \cup \{\langle y_1, y_2\rangle | y_1 \in Y_P, y_2 \in X_N\}$.

If $\dot{x} \in X_N$, it is output directly. For $\dot{x} \notin X_N$, i.e., $\dot{x} = \langle H, H(x)\rangle$, it outputs y according to the following probability density function (pdf):

$$pdf(\langle H, y\rangle | \langle H, H(x)\rangle) = \begin{cases} \frac{e^\varepsilon}{le^\varepsilon + 1 - l}, & \text{if } y \in C_{H(x)} \\ \frac{1}{le^\varepsilon + 1 - l}, & \text{if } y \in [0,1) \setminus C_{H(x)}, \end{cases} \tag{58}$$

where $C_{H(x)} = \{y | H(x) \le y < H(x) + l \text{ or } 0 \le y < H(x) + l - 1\}$.

For any $\langle H, y\rangle \in Y_P$, it is obvious that Eq. (15) can be maximized when $y \in C_{H(x_1)}$. At this point, $p_1 = \frac{e^\varepsilon}{le^\varepsilon + 1 - l}$ and $p_i = 1$ $(i = 2, 3, \ldots, s)$. Thus, we have:

$$\frac{\Pr[\mathcal{A}(x_1) = y]}{\Pr[\mathcal{A}(x_2) = y]} = e^\varepsilon \frac{1}{(1-z)(le^\varepsilon + (1-l))} = e^\varepsilon. \tag{59}$$

Therefore, the maximum value of z is $\frac{l(e^\varepsilon - 1)}{le^\varepsilon + 1 - l}$. If a non-sensitive data x is transformed into a sensitive data during the encoding process, it will have a probability of $\frac{l(e^\varepsilon - 1)}{le^\varepsilon + 1 - l}$ to additionally output x after the perturbation.

The Supp function for uWheel is defined as follows:

$$\text{Supp}(y) = \begin{cases} \{x | y \in C_{H(x)}\}, & \text{if } y = \langle H, y'\rangle \in Y_P \\ \{y\}, & \text{if } y \in X_N \\ \{x | y_1 \in C_{H(x)}\} \cup \{y_2\}, & \text{if } y = \langle \langle H, y_1\rangle, y_2\rangle. \end{cases} \tag{60}$$

Theorem 9. *Given the* Supp *function, the uWheel mechanism satisfies ULDP and is pure, with* $p^* = \frac{le^\varepsilon}{le^\varepsilon + 1 - l}$, $q^* = l$ *and* $z^* = \frac{(e^\varepsilon - 1)(sl - l + 1)}{(e^\varepsilon - 1)(s-1)l + (e^\varepsilon + s - 1)}$.

Aggregation. By substituting p^*, q^*, and z^* into Eq. (28), we can calculate the MSE of the estimated frequency of the uWheel mechanism.

$$MSE[\hat{\boldsymbol{c}}] = \frac{1}{n}\Big(s\frac{(e^\varepsilon l - l + 1)^2}{(e^\varepsilon - 1)^2(1-l)l} + (1-\theta)\frac{(l-1)^2 - e^\varepsilon l^2}{(e^\varepsilon - 1)(1-l)l} + \theta\frac{s}{(e^\varepsilon - 1)(ls - l + 1)}\Big). \tag{61}$$

Considering the application scenario of the uWheel mechanism, it is known that $s \gg \frac{1}{l} - 1$. To simplify the process of calculating partial derivatives and to avoid overly complex results, we let $MSE \approx \frac{1}{n}\big(s\frac{(e^\varepsilon l - l + 1)^2}{(e^\varepsilon - 1)^2(1-l)l} + (1-\theta)\frac{(l-1)^2 - e^\varepsilon l^2}{(e^\varepsilon - 1)(1-l)l} + \theta\frac{1}{(e^\varepsilon - 1)l}\big)$. Then, we take the partial derivative of the MSE.

$$\frac{\partial MSE[\hat{\boldsymbol{c}}]}{\partial g} = \frac{1}{(e^\varepsilon - 1)^2(l-1)^2 l^2 n}(e^{2\varepsilon} l^2(s + \theta - 1) - e^\varepsilon(l(l\theta - 2) + 1) - (l-1)^2(s-1)). \tag{62}$$

We denote the MSE as $m(l)$. Setting $m'(l) = 0$, we obtain two solutions and the only solution within $(0,1)$ is as follows:

$$l = \frac{1}{e^\varepsilon \sqrt{\frac{e^\varepsilon(s + \theta - 1) - \theta + 1}{e^\varepsilon(e^\varepsilon + s - 1)}} + 1}. \tag{63}$$

We observe that $\lim_{l \to 0^+} m'(l) < 0$ and $\lim_{l \to 1^-} m'(l) > 0$, indicating that Eq. (63) represents the optimal value for l.