

Providing Input-Discriminative Protection for Local Differential Privacy

Xiaolan Gu

*Department of ECE
University of Arizona
Tucson, AZ, USA*

xiaolang@email.arizona.edu

Ming Li

*Department of ECE
University of Arizona
Tucson, AZ, USA*

lim@email.arizona.edu

Li Xiong

*Department of Computer Science
Emory University
Atlanta, GA, USA*

lxiong@emory.edu

Yang Cao

*Department of Social Informatics
Kyoto University
Kyoto, Japan*

yang@i.kyoto-u.ac.jp

arXiv:1911.01402v2 [cs.CR] 24 Feb 2020

Abstract—Local Differential Privacy (LDP) provides provable privacy protection for data collection without the assumption of the trusted data server. In the real-world scenario, different data have different privacy requirements due to the distinct sensitivity levels. However, LDP provides the same protection for all data. In this paper, we tackle the challenge of providing input-discriminative protection to reflect the distinct privacy requirements of different inputs. We first present the Input-Discriminative LDP (ID-LDP) privacy notion and focus on a specific version termed MinID-LDP, which is shown to be a fine-grained version of LDP. Then, we focus on the application of frequency estimation and develop the IDUE mechanism based on Unary Encoding for single-item input and the extended mechanism IDUE-PS (with Padding-and-Sampling protocol) for item-set input. The results on both synthetic and real-world datasets validate the correctness of our theoretical analysis and show that the proposed mechanisms satisfying MinID-LDP have better utility than the state-of-the-art mechanisms satisfying LDP due to the input-discriminative protection.

Index Terms—local differential privacy, input-discriminative protection, frequency estimation

I. INTRODUCTION

Differential Privacy (DP) [1], [2] has become the *de facto* standard for private data release. It provides provable privacy protection, which is independent of the adversary’s background knowledge and computational power [3]. In recent years, Local Differential Privacy (LDP) has been proposed for preserving privacy at the data collection stage, in contrast to DP in the centralized setting which protects data after it is collected and stored by a server. In the local setting, the server is assumed to be untrusted, and each user randomly perturbs her raw data independently using a privacy-preserving mechanism that satisfies LDP. Then, the server collects these perturbed data from all users to perform data analytics or answer queries from users or third parties. Thus the local setting has been widely adopted in practice. For example, RAPPOR [4] proposed by Google has been employed in Chrome to collect web browsing behavior with LDP guarantees; Apple is also using LDP-based mechanism to identify popular emojis and popular health data types in Safari [5].

This work was partly supported by NSF grants CNS-1731164 and CNS-1618932, Air Force Office of Scientific Research (AFOSR) DDDAS program under grant FA9550-12-1-0240, JSPS KAKENHI grants with number 17H06099, 18H04093, 19K20269, and Microsoft Research Asia.

Under the notion of LDP, given any output of a mechanism, the adversary cannot distinguish any pair of inputs with high confidence (controlled by a privacy budget ϵ). Due to the uniform privacy budget, existing LDP mechanisms and applications [4], [6]–[8] would perturb the data in the same way (or add the noise with the same amount) for any inputs. However, in many practical scenarios, different inputs have different degrees of sensitivity (i.e., users’ desired privacy level or privacy expectation on the raw data) thus require distinct levels of privacy. For example, in website-click records or medical records, some website pages or medical diseases (e.g., HIV and cancer) are much more sensitive than others, thus need stronger privacy guarantees; on the other hand, some records are much less sensitive, such as commonly visited pages by many people (e.g., Facebook and Amazon), or some very common symptoms in clinic such as anemia and headache. Existing notions do not deal with this scenario. For example, personalized local differential privacy (PLDP) [3], [9] only provides user-level discrimination, and geoindistinguishability [10] only provides distance based discrimination for a pair of locations.

Motivated by such considerations, we consider the categorical data and assume the universe of inputs have multiple levels of privacy, represented by privacy budgets with different values. Note that a smaller budget indicates higher privacy requirement thus needs more protection. In practice, classifying items by privacy levels can be implemented according to some categories with semantic meanings. For example, serious diseases (e.g., various cancers or HIV) can be classified in the highest privacy level, while moderate diseases (e.g., asthma or hypertension) and common symptoms can be classified in the medium and lowest privacy levels respectively. Since each possible input x in domain \mathcal{D} has its privacy budget ϵ_x (inputs with the same privacy level have the same budget), the privacy budget of standard LDP should be $\epsilon = \min_{x \in \mathcal{D}} \{\epsilon_x\}$ to satisfy the required privacy for all inputs. Thus, LDP would provide excessive protection for some inputs that do not need such strong privacy, which is unnecessary and will lead to an inferior privacy-utility tradeoff.

In this paper, we aim at providing input-discriminative privacy with distinct protection for each input and high utility on frequency estimation. We first study how to formalize a privacy

notion in the local setting that provides discriminative privacy protection for different inputs. We propose a notion called Input-Discriminative LDP (ID-LDP) by converting the differentiated protection for inputs into different indistinguishability level for pairs of inputs. Theoretically, the indistinguishability of a pair of inputs x, x' can be any function of their budgets ϵ_x and $\epsilon_{x'}$. In this paper, we focus on one instantiation termed MinID-LDP with the minimum function. It relaxes LDP on the inputs that do not need too strong privacy protection, and we will show that the relaxation is at most twice of the minimum privacy budget of standard LDP (in Lemma 1). In summary, MinID-LDP can provide fine-grained protection where each input is protected with required indistinguishability, while LDP would overprotect the inputs that have less sensitivity.

Under our MinID-LDP notion, users need to perturb different inputs with different parameters related to the distinct privacy budgets, which makes the problem complicated since the perturbation parameters of a specific input may also depend on other inputs' privacy budgets to achieve indistinguishability between any two possible inputs. To find the optimal mechanism for a real-world query function, a potential solution is to formulate an optimization problem with the goal of maximizing query utility given privacy as constraints. However, the objective function of minimizing the Mean Squared Error (MSE) of the unbiased estimator is dependent on the unknown true frequencies thus cannot be directly evaluated. Also, the computation complexity is high because MinID-LDP considers multiple different privacy budgets, which leads to large numbers of variables (perturbation parameters need to be solved) and privacy constraints (which should be satisfied for any inputs x, x' and output y).

In this paper, we design two efficient and near-optimal mechanisms satisfying ID-LDP for frequency estimation on single-item and item-set data respectively. First, we propose Input-Discriminative Unary Encoding (IDUE) mechanism for *single-item input*. The objective function in optimization problem of assigning the perturbation probabilities in IDUE is approximated to be independent of the unknown true frequencies, and the number of variables and privacy constraints are $2t$ and t^2 respectively (t is the number of privacy levels). Note that the MSE of the naive mechanism without encoding (discussed in Sec. V-A) does not have closed-form expression and is dependent on the unknown true frequencies (thus the objective function cannot be directly evaluated), and the corresponding optimization problem has t^2 variables and t^3 constraints.

The proposed mechanism IDUE works well for single-item data. However, when the input is an *item-set*, i.e., any subset of the item domain, solving the optimization problem to determine the perturbation probabilities is not scalable due to an exponential blowup of the number of subsets. Thus, we combine our IDUE mechanism with Padding-and-Sampling protocol [7] to design a novel IDUE-PS mechanism for set-valued data. The privacy budget of a set is a function of the individual privacy budgets of items in the set. We will show that the perturbation probabilities of IDUE-PS (for item-set input with an exponential blowup) can be determined by IDUE

(for single-item input) to satisfy MinID-LDP (in Theorem 4) with a scalable optimization problem. Given the privacy level of each input, our proposed mechanisms satisfying MinID-LDP provide better privacy-utility tradeoff than ϵ -LDP (where $\epsilon = \min_{x \in \mathcal{D}} \{\epsilon_x\}$). It is because our mechanisms achieve fine-grained privacy protection; whereas, the existing mechanisms satisfying LDP guarantee the highest privacy level.

Main contributions are summarized as follows:

(1) We introduce a new privacy notion called Input-Discriminative LDP (ID-LDP) with an instantiation termed MinID-LDP, which allows finer-grained protection for different inputs than LDP.

(2) We design the Input-Discriminative Unary Encoding (IDUE) mechanism for single-item input that satisfies MinID-LDP and propose the frequency estimation protocol with an unbiased estimator. To minimize the Mean Squared Error (MSE) of IDUE, we formulate an optimization problem to solve the perturbation probabilities for the mechanism and derive three practical variants of the optimization model.

(3) We extend IDUE with Padding-and-Sampling into IDUE-PS for frequency estimation of item-set data, and show that it satisfies MinID-LDP with the same computation cost as IDUE that is designed for single-item input.

(4) We validate the correctness of the theoretical MSE analysis and effectiveness of our notion and mechanisms on synthetic and real-world datasets with both single-item and item-set types of input. We show that the proposed mechanisms outperform the existing ones for frequency estimation on categorical data. Also, the advantage of our mechanisms under the notion of MinID-LDP is enhanced when the distribution of privacy budgets of all inputs are more skewed.

II. RELATED WORK

The notion of differential privacy (DP) in centralized setting was first introduced by Dwork in [1]. It assumes a trusted server that possesses all genuine dataset. Then, a number of variants of differential privacy have been studied to provide different types of privacy guarantees such as d -privacy [11], Pufferfish privacy [12], Blowfish privacy [13], Concentrated DP [14], and Personalized DP [15]. On the other hand, Duchi et al. [16] studied local differential privacy (LDP) without the assumption of a trusted server, and many mechanisms are proposed to applied to diverse data types/tasks, such as frequency estimation [4], [6], set-valued data [7], and key-value data [8], [17]. Several variants of LDP and the corresponding mechanisms have been studied, e.g., Personalized LDP [3], [18], Geo-indistinguishability [10], [19], [20], Condensed LDP [21] and Utility-optimized LDP [22]. We will compare these notions with the proposed one in Sec. IV-B.

III. PROBLEM STATEMENT AND PRELIMINARIES

A. Problem Statement

System Model. Our system model involves one data server and n users $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$. Each user possesses one item or item-set in an item universe $\mathcal{I} = \{1, 2, \dots, m\}$ and

perturbs it independently via a random perturbation mechanism before uploading it to the server. Then, the server collects users' data and computes the statistical information of users' data (we focus on frequency estimation in this paper). We consider two types of input data, one is the single-item input with domain \mathcal{I} , where each user only possesses one item from \mathcal{I} ; another is the item-set input with domain $\mathcal{P}(\mathcal{I}) \triangleq \{x|x \subseteq \mathcal{I}\}$ (i.e., the power set of \mathcal{I} with size 2^m), where each user can possess any subset of \mathcal{I} . Assume there are t privacy levels, and the i -th level only contains a subset of \mathcal{I} , denoted by \mathcal{I}_i . Though the domain of the items can be large, the number of privacy levels determined by categories is usually small in practice, hence the usability and scalability of the system are guaranteed. For convenience, we denote the set of privacy budgets of all items in \mathcal{I} as $\mathcal{E} = \{\epsilon_i\}_{i \in \mathcal{I}}$.

Threat Model. We assume the server is untrusted, and each user only trusts herself, because data stored on the server can be revealed via either hacking activities or due to the server selling the data to a third party. Therefore, the adversary is assumed to possess the uploaded (perturbed) data of all users and it also knows the perturbation mechanism and the privacy budgets for all the inputs.

Utility of Frequency Estimation. The true frequency of an item $i \in \mathcal{I}$ is defined as the number of users who possess i

$$c_i^* = \sum_{u \in \mathcal{U}} \mathbb{1}_{x_u}(i) \quad (\forall i \in \mathcal{I}) \quad (1)$$

where x_u is the raw (input) data of a user $u \in \mathcal{U}$ and can be a single-item or an item-set depending on the application scenario, and $\mathbb{1}_{x_u}(i)$ is the indicator function, which is equal to 1 if $i \in x_u$ and equal to 0 otherwise. Note that i only denotes one item from \mathcal{I} , while x_u can be a subset of \mathcal{I} . After collecting the perturbed (output) data from all users, the server can estimate the frequency of an item $i \in \mathcal{I}$ via an estimator \hat{c}_i , which is a function of the perturbed data $\{y_u\}_{u \in \mathcal{U}}$ and mechanism parameters. The utility of frequency estimation is defined by the total Mean Squared Error (MSE) of estimators, i.e., $\text{MSE} = \sum_{i=1}^m \text{MSE}_{\hat{c}_i}$, which will be minimized in the design of mechanism with privacy constraints.

B. The Notion of LDP

In the local setting, each user independently perturbs her input x (raw data) using a mechanism \mathcal{M} and uploads $\mathcal{M}(x)$ to the server for data analysis.

Definition 1 (Local Differential Privacy (LDP) [16]) For a given $\epsilon \in \mathbb{R}^+$, a randomized mechanism \mathcal{M} satisfies ϵ -LDP if and only if for any pair of inputs x, x' and any output y

$$\frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} \leq e^\epsilon \quad (2)$$

Intuitively, given an output y of a mechanism \mathcal{M} , an adversary cannot infer with high confidence (controlled by ϵ) whether the input is x or x' , which provides plausible deniability for individuals involved in the sensitive data. Here, ϵ is a parameter called *privacy budget* that controls the strength of privacy protection. A smaller ϵ indicates stronger privacy protection because the adversary has lower confidence

when trying to distinguish any pair of inputs x, x' . LDP has the property of sequential composition, which guarantees the overall privacy for a sequence of mechanisms that satisfy LDP. **Theorem 1 (Sequential Composition of LDP [23])** If randomized mechanism $\mathcal{M}_i : \mathcal{D} \rightarrow \mathcal{R}_i$ satisfies ϵ_i -LDP for $i = 1, 2, \dots, k$, then their sequential combination $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}_1 \times \mathcal{R}_2 \times \dots \times \mathcal{R}_k$ defined by $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k)$ satisfies $(\sum_{i=1}^k \epsilon_i)$ -LDP.

According to sequential composition, a given privacy budget ϵ can be split into multiple portions, where each portion corresponds to the privacy budget of a randomized mechanism.

C. Mechanisms Satisfying LDP

Randomized Response. Randomized Response (RR) [24] is a technique developed for the participants in a survey to return a randomized answer to a sensitive question to protect their privacy. Specifically, each participant gives a genuine answer with probability p or gives the opposite answer with probability $1 - p$, where $p = \frac{e^\epsilon}{e^\epsilon + 1}$ to satisfy ϵ -LDP. The standard RR only works for binary data (yes-or-no answers), but it can be extended to apply to m categories by Generalized Randomized Response or Unary Encoding.

Generalized Randomized Response. The perturbation function in Generalized Randomized Response (GRR) [7] is

$$\Pr(\mathcal{M}(x) = y) = \begin{cases} p, & \text{if } y = x \\ q, & \text{if } y \neq x \end{cases}, \quad (\forall x, y = 1, 2, \dots, m)$$

To satisfy ϵ -LDP, the probabilities are $p = \frac{e^\epsilon}{e^\epsilon + m - 1}$ and $q = \frac{1}{e^\epsilon + m - 1}$, both of which would be small when the domain size m is very large compared with e^ϵ .

Unary Encoding. Unary Encoding (UE) [6] converts the input $x = i$ into a vector $\mathbf{x} = [0, \dots, 0, 1, 0, \dots, 0]$ with length m where only the i -th bit is 1. Then each user perturbs each bit of \mathbf{x} independently with the following probabilities

$$\Pr(\mathbf{y}[k] = 1) = \begin{cases} p, & \text{if } \mathbf{x}[k] = 1 \\ q, & \text{if } \mathbf{x}[k] = 0 \end{cases} \quad (\forall k = 1, 2, \dots, m)$$

This mechanism satisfies LDP with $\epsilon = \ln \frac{p(1-q)}{(1-p)q}$ [6]. The selection of p and q under a given privacy budget ϵ varies for different mechanisms. For example, the basic RAPPOR [4] assigns $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1}$, $q = 1 - p$, while the Optimized Unary Encoding (OUE) [6] assigns $p = 0.5$, $q = \frac{1}{e^\epsilon + 1}$, which are obtained by optimizing the approximate variance.

Frequency Estimation for GRR, RAPPOR and OUE. After receiving the perturbed data from all users, the server can implement the summation to get the total count of each bit, denoted by c_i for the i -th bit. Then, the server calibrates the collected counts by an unbiased estimator $\hat{c}_i = \frac{c_i - nq}{p - q}$, whose Mean Squared Error (MSE) is equal to its variance [6]

$$\text{MSE}_{\hat{c}_i} = \text{Var}[\hat{c}_i] = \frac{nq(1-q)}{(p-q)^2} + \frac{c_i^*(1-p-q)}{p-q}$$

where c_i^* is the ground truth of the counting for item i . In summary, OUE can provide higher utility than RAPPOR for frequency estimation under the same ϵ due to the optimization,

and the utility of GRR would be deteriorated much more than the other two mechanisms when the domain size m is large.

IV. INPUT-DISCRIMINATIVE LDP

In this section, a new privacy notion called ID-LDP is introduced, which can provide input-discriminative protection with LDP. In ID-LDP, the indistinguishability level of a pair of possible inputs x, x' is determined by the corresponding privacy budgets $\epsilon_x, \epsilon_{x'}$ of both inputs. Then, one instantiation of ID-LDP called MinID-LDP is formalized. It is proven to satisfy sequential composition theorem, which is an important property to guarantee the overall privacy for multiple query functions sequentially applied to the same database. Finally, our notion is compared with several existing privacy notions in the local setting and their relations are discussed.

A. Definition

LDP defines privacy as the maximum level of indistinguishability between any two possible inputs. In practical applications, the privacy levels of different inputs could be distinct. Thus, the requirement of indistinguishability between different pairs of inputs could be diverse. However, LDP cannot provide such fine-grained privacy protection because its definition is based on the worst-case scenario. Intuitively, discriminating inputs with different privacy levels and providing distinct protection to them can improve the utility of the query service due to the fine-grained protection for different inputs. We define the new notion ID-LDP as follows.

Definition 2 (Input-Discriminative LDP (ID-LDP)) For a given privacy budget set $\mathcal{E} = \{\epsilon_x\}_{x \in \mathcal{D}} \in \mathbb{R}_+^{|\mathcal{D}|}$, where $|\mathcal{D}|$ is the size of the input domain \mathcal{D} , the randomized mechanism \mathcal{M} satisfies \mathcal{E} -ID-LDP if and only if for any pair of inputs $x, x' \in \mathcal{D}$, and any output $y \in \text{Range}(\mathcal{M})$

$$\frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} \leq e^{r(\epsilon_x, \epsilon_{x'})} \quad (3)$$

where $r(\cdot, \cdot)$ is a function of two privacy budgets.

In Definition 2, we assume inputs x and x' belong to different privacy levels with privacy budgets ϵ_x and $\epsilon_{x'}$ respectively and introduce a system-defined function $r(\epsilon_x, \epsilon_{x'})$ to quantify the indistinguishability between x and x' . Note that the value of ϵ_x for each input x is not sensitive information because ϵ_x is independent of the users' raw data. In this paper, we assume $\{\epsilon_x\}_{x \in \mathcal{D}}$ are universally set by the service provider. Note that, our notion can be easily combined with personalized LDP (PLDP) to reflect different privacy preferences of different users, in which case the privacy levels of all inputs can be set by users themselves. Theoretically, the notion of ID-LDP does not restrict the data type, which means it can be applied for categorical data, numerical data, or even the hybrid with multi-dimensions. In this paper, we mainly study the mechanism that satisfies ID-LDP for categorical data (single-item or item-set).

ID-LDP can provide input-discriminative protection with the function $r(\cdot, \cdot)$. In this paper, we mainly consider the minimum function between ϵ_x and $\epsilon_{x'}$ as the privacy budget of a pair of inputs x, x' , formulated by the following definition.

Definition 3 (MinID-LDP) A randomized mechanism \mathcal{M} satisfies \mathcal{E} -MinID-LDP if and only if it satisfies \mathcal{E} -ID-LDP with $r(\epsilon_x, \epsilon_{x'}) = \min\{\epsilon_x, \epsilon_{x'}\}$.

Intuitively, for any pair of inputs x, x' , MinID-LDP guarantees that the adversary's capability of distinguishing x and x' would not exceed the bound controlled by both ϵ_x and $\epsilon_{x'}$, which achieves the worse-case privacy like LDP but *only for the pair*. We use an example to show the benefit of our notion.

Example. Assume a health organization is taking a survey which asks n participants to return a response perturbed from categories {HIV, anemia, headache, stomachache, toothache}, indexed by an integer i from {1,2,3,4,5}. Since HIV ($i = 1$) is more sensitive than the others, the privacy budget that represents the privacy level should be different, such as $\epsilon_1 = \ln 4$ for HIV and $\epsilon_i = \ln 6$ ($i \neq 1$) for the others, where a smaller ϵ indicates a higher privacy level that needs stronger privacy protection. To satisfy LDP, all categories will be perturbed under the privacy budget $\epsilon_1 = \ln 4$, even though some of them (such as anemia and headache) do not need such strong privacy protection. Under MinID-LDP, however, anemia and headache can be perturbed with less noise as long as the indistinguishability of any pair of inputs is bounded by both two budgets of them. We will compare the utility of mechanisms under the two notions in Sec. V-E.

As mentioned in Sec. III-B, sequential composition is an important property to guarantee the overall privacy for a sequence of mechanisms. The following theorem shows that MinID-LDP satisfies sequential composition as well.

Theorem 2 (Sequential Composition of MinID-LDP) If randomized mechanism $\mathcal{M}_i : \mathcal{D} \rightarrow \mathcal{R}_i$ satisfies \mathcal{E}_i -MinID-LDP for $i = 1, 2, \dots, k$, where $\mathcal{E}_i = \{\epsilon_x^{(i)}\}_{x \in \mathcal{D}} \in \mathbb{R}_+^{|\mathcal{D}|}$, then their combination $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}_1 \times \mathcal{R}_2 \times \dots \times \mathcal{R}_k$ defined by $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k)$ satisfies $(\sum_{i=1}^k \mathcal{E}_i)$ -MinID-LDP, where $(\sum_{i=1}^k \mathcal{E}_i) \triangleq \{\sum_{i=1}^k \epsilon_x^{(i)}\}_{x \in \mathcal{D}}$.

Proof: Let $x, x' \in \mathcal{D}$ be any pair of inputs, for any output $y = (y_1, y_2, \dots, y_k) \in \mathcal{R}_1 \times \mathcal{R}_2 \times \dots \times \mathcal{R}_k$, we have

$$\begin{aligned} \frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} &= \prod_{i=1}^k \frac{\Pr(\mathcal{M}_i(x) = y_i)}{\Pr(\mathcal{M}_i(x') = y_i)} \leq \prod_{i=1}^k e^{\min\{\epsilon_x^{(i)}, \epsilon_{x'}^{(i)}\}} \\ &\leq \prod_{i=1}^k e^{\epsilon_x^{(i)}} = e^{\sum_{i=1}^k \epsilon_x^{(i)}} \end{aligned}$$

Similarly, $\frac{\Pr(\mathcal{M}(x)=y)}{\Pr(\mathcal{M}(x')=y)} \leq e^{\sum_{i=1}^k \epsilon_{x'}^{(i)}}$. Finally, we have

$$\frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} \leq e^{\min\{\sum_{i=1}^k \epsilon_x^{(i)}, \sum_{i=1}^k \epsilon_{x'}^{(i)}\}}$$

which indicates that \mathcal{M} satisfies $(\sum_{i=1}^k \mathcal{E}_i)$ -MinID-LDP. ■

B. Relationships and Comparison with Other Notions

Relationships with LDP. If the privacy budgets for all inputs are the same, i.e., $\epsilon_x = \epsilon$ for all $x \in \mathcal{D}$, then \mathcal{E} -MinID-LDP becomes ϵ -LDP, which means MinID-LDP is a generalized version of LDP. In general, we have the following lemma to show their relationships.

Lemma 1 If a mechanism satisfies ϵ -LDP, then it also satisfies \mathcal{E} -MinID-LDP for all \mathcal{E} with $\min\{\mathcal{E}\} = \epsilon$. On the other hand,

if a mechanism satisfies \mathcal{E} -MinID-LDP, then it also satisfies ϵ -LDP, where $\epsilon = \min\{\max\{\mathcal{E}\}, 2 \min\{\mathcal{E}\}\}$.

Proof: First, the following property can be directly derived from the definitions of LDP and MinID-LDP

$$\min\{\mathcal{E}\}\text{-LDP} \Rightarrow \mathcal{E}\text{-MinID-LDP} \Rightarrow \max\{\mathcal{E}\}\text{-LDP}$$

Therefore, we only need to show that \mathcal{E} -MinID-LDP also implies $2 \min\{\mathcal{E}\}$ -LDP. Denote x^* as the input that has the minimum budget, i.e., $\epsilon_{x^*} = \min\{\mathcal{E}\}$. Then, for all x, x' and y , the following inequality is satisfied under \mathcal{E} -MinID-LDP

$$\begin{aligned} \frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} &= \frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x^*) = y)} \cdot \frac{\Pr(\mathcal{M}(x^*) = y)}{\Pr(\mathcal{M}(x') = y)} \\ &\leq e^{\epsilon_{x^*}} \cdot e^{\epsilon_{x^*}} = e^{2\epsilon_{x^*}} = e^{2\min\{\mathcal{E}\}} \end{aligned}$$

which means \mathcal{E} -MinID-LDP implies $2 \min\{\mathcal{E}\}$ -LDP. ■

From Lemma 1, MinID-LDP relaxes LDP in at most twice of the privacy budget $\epsilon = \min\{\mathcal{E}\}$. It is due to the symmetric property of the indistinguishability definition, so in a fully-connected policy graph, if we require every pair of inputs x, x' to be indistinguishable with $\min\{\epsilon_x, \epsilon_{x'}\}$, transitivity of indistinguishability yields $2 \min\{\mathcal{E}\}$ between any pair of inputs. Note that the twice relaxation in privacy budget does not mean utility improvement is at most twice compared to LDP (depending on the query and data distribution). Although MinID-LDP can be regarded as a relaxation compared with LDP, in practice users' privacy expectation is naturally different for different inputs, hence our notion captures user's fine-grained requirement, while LDP is too strong (i.e., provides overprotection) in this regard.

Related Privacy Notions. LDP provides the worst-case privacy protection for all users and all inputs, where the global privacy budget is $\epsilon = \min_{x \in \mathcal{D}}\{\epsilon_x\}$. Several variants of LDP are related to our notion, but they have different ideas. Fig. 1 shows the differences of personalized LDP (PLDP) [18], geo-indistinguishability (GI) [10], condensed LDP (CLDP) [21], and our notion ID-LDP, in the form of their privacy policies where the vertices are inputs and edges are the distinguishability level (represented by privacy budget) of each pair of inputs. PLDP in [9], [18] provides user-discriminative privacy requirements, i.e., each user can have personalized privacy budget which is often assumed to be unrelated to the raw data if it would be published. Another PLDP notion [3] considers both safe region and privacy budget for each user in location-based systems. In summary, PLDP provides different protections for different users but does not differentiate different pairs of inputs. On the other hand, geo-indistinguishability [10] in location-based systems and CLDP [21] in the data collection setting can provide distance-discriminative privacy, but they originate from an input pair-centric viewpoint and requires a distance metric for the inputs, where the distance metric (satisfying triangle inequality) may be hard to define for some data types such as categorical data. In contrast, our notion ID-LDP provides input-discriminative privacy requirements, where each input has a privacy budget (inputs with the same privacy level have the same budget), and

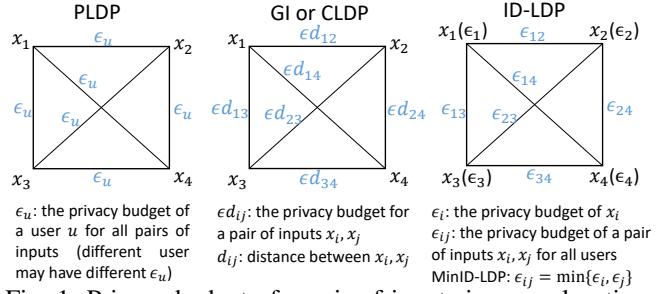


Fig. 1: Privacy budget of a pair of inputs in several notions.

TABLE I: The bounds of prior-posterior $\frac{\Pr(x)}{\Pr(x|y)}$ ($\forall x, y$).

Privacy Notions	Lower Bound	Upper Bound
LDP	$e^{-\epsilon}$	e^ϵ
PLDP	$e^{-\epsilon_u}$	e^{ϵ_u}
GI or CLDP	$\sum_{x'} \Pr(x') e^{-\epsilon \cdot d(x, x')}$	$\sum_{x'} \Pr(x') e^{\epsilon \cdot d(x, x')}$
MinID-LDP	$e^{-\min\{\epsilon_x, 2 \min\{\mathcal{E}\}\}}$	$e^{\min\{\epsilon_x, 2 \min\{\mathcal{E}\}\}}$

the distinguishability of a pair of inputs can be determined by a function of the budgets of the two inputs to bound the distinguishability of this pair. Another notion that also considers distinct privacy levels is Utility-optimized LDP (ULDP) [22], which provides a privacy guarantee equivalent to LDP only for sensitive data to add less noise and improve utility. It can be regarded as a special case of the proposed MinID-LDP under two privacy levels of inputs (sensitive and non-sensitive) but with incomplete privacy policy graph, where sensitive and non-sensitive inputs can be fully distinguished when observing some outputs (termed invertible data) that reveals non-sensitive inputs, thus ULDP does not guarantee LDP. However, our MinID-LDP relaxes LDP by providing distinct bounds of privacy leakage for multiple (more than two) privacy levels of inputs and also guarantees LDP with some privacy budgets (refer to Lemma 1).

Prior-Posterior Privacy Leakage Analysis. To understand our privacy notions in another perspective, we compare the prior-posterior privacy leakage (i.e., Local Information Privacy [25]) of the above notions. Denote $\Pr(y|x)$ as the probability of outputting y by given input x . The ratio between the prior probability $\Pr(x)$ of an input x and the posterior probability $\Pr(x|y)$ by observing the output y can be computed as

$$\frac{\Pr(x)}{\Pr(x|y)} = \frac{\Pr(y)}{\Pr(y|x)} = \frac{\sum_{x' \in \mathcal{D}} \Pr(x') \Pr(y|x')}{\Pr(y|x)} \quad (4)$$

which quantifies the privacy leakage that the additional information the adversary can infer about an input x by observing the output y . Note that (4) is different from Mutual Information (MI) [26] that quantifies the average leakage for all inputs and outputs. In our case, we only evaluate the bound of privacy leakage for a given input x with an arbitrary output y . For different privacy notions, the lower bound and upper bound (independent of y) of prior-posterior privacy leakage defined by (4) are summarized in Table I (can be directly derived from definitions or Lemma 1). The notion of LDP, PLDP, and MinID-LDP have the similar bound of privacy leakage for input x with respect to the budget (though MinID-LDP

has an additional bound with respect to $2 \min\{\mathcal{E}\}$). However, LDP and PLDP do not differentiate the inputs, thus the budget would be assigned as the minimum value of all budgets in order to satisfy the privacy, but MinID-LDP can assign the required budget for different inputs, where the bound of leakage is also input-discriminative.

V. PERTURBATION MECHANISM AND FREQUENCY ESTIMATION FOR SINGLE-ITEM INPUT

In this section, the considered input domain is $\mathcal{D} = \mathcal{I}$, i.e., single-item input. First, we formulate the optimization problem for designing a perturbation mechanism to optimize the utility of the frequency estimation of the outputs while satisfying MinID-LDP and the challenges to solve the problem. To address the challenges, we propose Input-Discriminative Unary Encoding (IDUE) mechanism and the corresponding unbiased frequency estimator. Finally, we develop three practical variants of optimization model to obtain the optimal (or near-optimal) perturbation probabilities in IDUE.

A. Objectives and Challenges

Our goal is to design a framework with perturbation mechanism and frequency estimation protocol that satisfies the proposed notion ID-LDP (MinID-LDP specifically) with the optimal Mean Squared Error (MSE) of frequency estimation. The general optimization problem can be modeled as

$$\min \text{ MSE}, \quad s.t. \quad \frac{\Pr(y|x)}{\Pr(y|x')} \leq e^{r(\epsilon_x, \epsilon_{x'})} \quad (\forall x, x', y)$$

However, solving this optimization problem has two challenges. First, the objective function cannot be directly evaluated in general because MSE is dependent on the unknown true frequencies. Second, the computation complexity is high since the numbers of variables (perturbation parameters/probabilities that determine the ratio $\frac{\Pr(y|x)}{\Pr(y|x')}$ need to be solved) and privacy constraints (which should be satisfied for any inputs x, x' and output y) can be very large.

For example, a direct way to design such mechanism is to assign a perturbation/mapping matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$ (which can be regarded as a variant of GRR discussed in Sec. III-C), where each element represents the perturbation probability $\Pr(y|x)$ for $x, y \in \mathcal{D}$ (the output domain $\mathcal{R} = \mathcal{D}$ in this case). However, solving the elements in matrix \mathbf{P} by minimizing MSE under the privacy constraints has several issues in practice. First, $\hat{\mathbf{c}} = (\mathbf{P}^T)^{-1}\mathbf{c}$ was shown to be the unbiased estimator of the true frequency vector \mathbf{c}^* [27], where \mathbf{c} is the calculated frequency vector of outputs. But the elements in the inversion matrix $(\mathbf{P}^T)^{-1}$ do not have closed-form expression in general and the MSE of this estimation is dependent on the unknown true frequencies, thus the objective function of minimizing MSE cannot be directly evaluated. Note that the frequency estimator of the original GRR discussed in Sec. III-C can be regarded as a special case of the above mechanism, where the inversion matrix can be explicitly calculated (because there are only two different perturbation probabilities) and the term that is related to the

true frequencies only takes a small portion of MSE, thus the *approximate* MSE can be independent of the unknown true frequencies. But it cannot be applied here because the perturbation probabilities for different inputs are designed to be different in our setting. Second, since the numbers of variables and constraints are $|\mathcal{D}|^2$ (all elements in \mathbf{P}) and $|\mathcal{D}|^3$ (for all x, x' and y) respectively, the computation cost would be very high especially for item-set input $\mathcal{D} = \mathcal{P}(\mathcal{I})$ with $|\mathcal{D}| = 2^m$. Third, the domain size $|\mathcal{D}|$ is usually very large in practice, then the perturbation probabilities will become very small because of $\sum_{y \in \mathcal{D}} \Pr(y|x) = 1$ for all $x \in \mathcal{D}$, which means the probability of reporting the true value is low, then the utility would greatly deteriorate.

In the following, we propose the Unary Encoding (refer to Sec. III-C) based perturbation mechanism and frequency estimation protocol for single-item input with $\mathcal{D} = \mathcal{I}$ under privacy notion ID-LDP. Due to the nature of Unary Encoding, there are less perturbation parameters (shown in Sec. V-B), and the upper bound for all y of ratio $\frac{\Pr(y|x)}{\Pr(y|x')}$ when fixing x and x' can be explicitly calculated. Then, the equivalent constraint in (6) needs to be satisfied only for all inputs x and x' (hence the number of constraints is reduced). On the other hand, the unbiased estimator \hat{c}_i in (7) can be explicitly expressed by the perturbation probabilities, and its MSE in (8) can be composed by two terms, where only the second term is dependent on the true frequency c_i^* . Finally, we address the challenge due to the lack of true frequencies by developing three variants of optimization models in Sec. V-D to obtain the approximate total MSE which is independent of true frequencies.

B. Mechanism Design

Input-Discriminative Unary Encoding (IDUE). We first encode the single-item input $x = i$ into a m -length vector

$$\mathbf{x} = \mathbf{v}_i = [0, \dots, 0, 1, 0, \dots, 0] \quad (5)$$

where vector \mathbf{x} denotes the encoded input, \mathbf{v}_i denotes the vector whose i -th position is 1 and other positions are 0s. Then, each bit of the input vector \mathbf{x} is perturbed into 0 or 1 independently to get the output vector \mathbf{y} with probabilities

$$\begin{aligned} \Pr(\mathbf{y}[k] = 1 | \mathbf{x}[k] = 1) &= a_k, \quad \Pr(\mathbf{y}[k] = 0 | \mathbf{x}[k] = 1) = 1 - a_k \\ \Pr(\mathbf{y}[k] = 1 | \mathbf{x}[k] = 0) &= b_k, \quad \Pr(\mathbf{y}[k] = 0 | \mathbf{x}[k] = 0) = 1 - b_k \end{aligned}$$

where we assume $a_k > b_k$ ($\forall k \in \mathcal{I}$) in order to obtain a good utility. Compared with the original Unary Encoding protocol [6] discussed in Sec. III-C, where p and q correspond to a_i and b_i in our notation, IDUE assigns different perturbation probabilities for different bits, which is the key point to achieve input-discriminative protection.

For two different input vectors \mathbf{v}_i (only the i -th bit is 1) and \mathbf{v}_j , the probability ratio of distinguishing the pair of \mathbf{v}_i and \mathbf{v}_j by observing the output vector \mathbf{y} is

$$\frac{\Pr(\mathbf{y}|\mathbf{v}_i)}{\Pr(\mathbf{y}|\mathbf{v}_j)} = \frac{\prod_{k=1}^m \Pr(\mathbf{y}[k]|\mathbf{v}_i)}{\prod_{k=1}^m \Pr(\mathbf{y}[k]|\mathbf{v}_j)} = \frac{\Pr(\mathbf{y}[i]|\mathbf{v}_i) \Pr(\mathbf{y}[j]|\mathbf{v}_i)}{\Pr(\mathbf{y}[i]|\mathbf{v}_j) \Pr(\mathbf{y}[j]|\mathbf{v}_j)}$$

Since $a_k > b_k$ ($\forall k \in \mathcal{I}$), we have

$$\frac{\Pr(\mathbf{y}[i]|\mathbf{v}_i)\Pr(\mathbf{y}[j]|\mathbf{v}_i)}{\Pr(\mathbf{y}[i]|\mathbf{v}_j)\Pr(\mathbf{y}[j]|\mathbf{v}_j)} = \frac{\left(\frac{a_i}{b_i}\right)^{\mathbf{y}[i]} \left(\frac{1-a_i}{1-b_i}\right)^{1-\mathbf{y}[i]}}{\left(\frac{a_j}{b_j}\right)^{\mathbf{y}[j]} \left(\frac{1-a_j}{1-b_j}\right)^{1-\mathbf{y}[j]}} \leq \frac{a_i(1-b_j)}{b_i(1-a_j)}$$

where the left side equals the right side if and only if $\mathbf{y}[i] = 1$ and $\mathbf{y}[j] = 0$. Then, the privacy constraint in (3) is

$$\frac{a_i(1-b_j)}{b_i(1-a_j)} \leq e^{r(\epsilon_i, \epsilon_j)} \quad (\forall i, j \in \mathcal{I}) \quad (6)$$

By converting the original privacy constraint into (6), which is independent of y thus has less number of constraints, we can reduce the computational complexity compared with the direct formulation described in Sec. V-A.

To obtain the optimal perturbation probabilities for our IDUE mechanism, we first develop the frequency estimator for IDUE, and evaluate the theoretical MSE of the estimator as a function of perturbation probabilities. Then we formalize the optimization problem with three variants by minimizing the MSE with the privacy constraints in (6).

C. Frequency Estimation

Denote the collected frequency of the i -th bit as $c_i = \sum_{u \in \mathcal{U}} \mathbf{y}_u[i]$, where \mathbf{y}_u is the output vector of a user $u \in \mathcal{U}$. For frequency estimation, we utilize the following estimator

$$\hat{c}_i = \frac{c_i - nb_i}{a_i - b_i} \quad (7)$$

which can be shown as the unbiased estimator of the true frequency c_i^* defined in (1).

Theorem 3 (Unbiasedness Property) If $a_i \neq b_i$ ($\forall i \in \mathcal{I}$), then $\mathbb{E}[\hat{c}_i] = c_i^*$, where estimator \hat{c}_i is defined in (7).

Proof: Since $\mathbb{E}[c_i] = c_i^* a_i + \sum_{k \neq i} c_k^* b_i = c_i^* a_i + (n - c_i^*) b_i$, then we have $\mathbb{E}[\hat{c}_i] = \frac{\mathbb{E}[c_i] - nb_i}{a_i - b_i} = c_i^*$, which means \hat{c}_i is an unbiased estimator of c_i^* . ■

The frequency estimator in (7) can be regarded as the generalized version of the estimator that is used for the original Unary Encoding (refer to Sec. III-C). Due to the unbiasedness of estimator \hat{c}_i , the MSE of \hat{c}_i is equal to its variance

$$\begin{aligned} \text{MSE}_{\hat{c}_i} &= \text{Var}[\hat{c}_i] = \frac{c_i^* a_i (1 - a_i) + (n - c_i^*) b_i (1 - b_i)}{(a_i - b_i)^2} \\ &= \frac{nb_i(1 - b_i)}{(a_i - b_i)^2} + \frac{c_i^*(1 - a_i - b_i)}{a_i - b_i} \end{aligned} \quad (8)$$

In Sec. V-D, the summation of $\text{MSE}_{\hat{c}_i}$ will be minimized with the privacy constraints of ID-LDP.

D. Finding Optimal Perturbation Probabilities

As described in Sec. III-A, the input domain is divided into t subsets $\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_t$ with different privacy levels. Denote the number of items in subset \mathcal{I}_i as $|\mathcal{I}_i| = m_i$ and the privacy budget is ϵ_i ($i = 1, 2, \dots, t$). We can assign the same parameters a_i and b_i for all items in \mathcal{I}_i . If $t = 1$, i.e., all items in \mathcal{I} have the same ϵ , then this case reduces to the LDP setting. The MSE of subset \mathcal{I}_i is calculated by

$$\text{MSE}_{\mathcal{I}_i} = \sum_{k \in \mathcal{I}_i} \text{MSE}_{\hat{c}_k} = \frac{nm_i b_i (1 - b_i)}{(a_i - b_i)^2} + \frac{(1 - a_i - b_i)}{a_i - b_i} \sum_{k \in \mathcal{I}_i} c_k^*$$

The expression of $\text{MSE}_{\mathcal{I}_i}$ is dependent on the true frequency $\sum_{k \in \mathcal{I}_i} c_k^*$, which is unknown in practice, thus cannot be established as the objective function for the optimization problem. Therefore, we propose three variants of the optimization model, named opt0, opt1, and opt2, to make the objective function independent of the true frequencies.

opt0: Optimization Model in the Worst-Case. Though $\text{MSE}_{\mathcal{I}_i}$ is dependent on the true frequencies, we have the following upper bound of the total MSE since $\sum_{k \in \mathcal{I}_i} c_k^* \leq n$ to get rid of the unknown true frequency c_k^*

$$\sum_{i=1}^t \text{MSE}_{\mathcal{I}_i} \leq \sum_{i=1}^t \frac{nm_i b_i (1 - b_i)}{(a_i - b_i)^2} + \max \left\{ \frac{1 - a_i - b_i}{a_i - b_i} \right\} \cdot n$$

which can be regarded as the total MSE in the worst-case. Then, determining parameters $\{a_i, b_i\}_{i=1}^t$ is converted to minimizing the worst-case MSE

$$\begin{aligned} \min_{a,b} \quad & f \triangleq \sum_{i=1}^t \frac{m_i b_i (1 - b_i)}{(a_i - b_i)^2} + \max \left\{ \frac{1 - a_i - b_i}{a_i - b_i} \right\} \\ \text{s.t.} \quad & \frac{a_i(1 - b_j)}{b_i(1 - a_j)} \leq e^{r(\epsilon_i, \epsilon_j)} \quad (\forall i, j = 1, 2, \dots, t) \\ & 0 < b_i < a_i < 1 \quad (\forall i = 1, 2, \dots, t) \end{aligned} \quad (9)$$

where the scaling constant n in the objective function is omitted since it does not change the result. Since the feasible region of optimization problem (9) contains the perturbation probabilities of RAPPOR and OUE, the solution solved by (9) will have less worst-case MSE than both RAPPOR and OUE.

It can be shown that the objective function in (9) is not convex in the feasible region. To address this, in the following we consider two types of space reducing strategies, which are related to RAPPOR and OUE respectively. They can be used to find near-optimal solutions with convex property and reduced complexity compared with the formulation in (9). Our idea is to further constrain the variables (which shrinks the feasible region), so that the optimization problem can be convex.

opt1: Optimization Model Constrained with RAPPOR Structure.

RAPPOR regards bit-0 and bit-1 equally thus $p + q = 1$. We add the corresponding constraint $a_i + b_i = 1$ ($\forall i$) in our optimization problem and represent a_i, b_i as

$$a_i = \frac{e^{\tau_i}}{e^{\tau_i} + 1}, \quad b_i = \frac{1}{e^{\tau_i} + 1} \quad (i = 1, 2, \dots, t) \quad (10)$$

where $\tau_i > 0$ ($\forall i$). Then $\frac{1-a_i-b_i}{a_i-b_i} = 0$ and the total MSE is

$$\sum_{i=1}^t \text{MSE}_{\mathcal{I}_i} = \sum_{i=1}^t \frac{nm_i b_i (1 - b_i)}{(a_i - b_i)^2} = n \sum_{i=1}^t \frac{m_i e^{\tau_i}}{(e^{\tau_i} - 1)^2}$$

with privacy constraints

$$\frac{a_i(1 - b_j)}{b_i(1 - a_j)} = e^{\tau_i + \tau_j} \leq e^{r(\epsilon_i, \epsilon_j)} \Leftrightarrow \tau_i + \tau_j \leq r(\epsilon_i, \epsilon_j)$$

TABLE II: Utility comparison in the toy example, where $\epsilon_1 = \ln 4$ and $\epsilon_i = \ln 6$ ($i \neq 1$).

Mechanisms	Privacy Notions	Probability of flipping the i -th bit				Variance of frequency estimation		$\sum_i \text{Var}[\hat{c}_i]$	
		1 - a_i (if $\mathbf{x}[i] = 1$)		b_i (if $\mathbf{x}[i] = 0$)		$\text{Var}[\hat{c}_i]$			
		$i = 1$	$i = 2 \sim 5$	$i = 1$	$i = 2 \sim 5$	$i = 1$	$i = 2 \sim 5$		
RAPPOR [4]	LDP	0.33	0.33	0.33	0.33	$2n$	$2n$	$10n$	
OUE [6]	LDP	0.5	0.5	0.2	0.2	$1.78n + c_i$	$1.78n + c_i$	$9.9n$	
IDUE	MinID-LDP	0.41	0.33	0.33	0.28	$3.27n + 0.31c_i$	$1.32n + 0.13c_i$	$8.68n \sim 8.86n$	

Therefore, we can get the following optimization problem

$$\begin{aligned} \min_{\tau_1, \dots, \tau_t > 0} \quad f(\tau) &\triangleq \sum_{i=1}^t \frac{m_i e^{\tau_i}}{(e^{\tau_i} - 1)^2} \\ \text{s.t.} \quad \tau_i + \tau_j &\leq r(\epsilon_i, \epsilon_j) \quad (\forall i, j) \end{aligned} \quad (11)$$

opt2: Optimization Model Constrained with OUE Structure. OUE focuses on less noise of bit-0 thus $p = 0.5$. We add the additional constraints $a_i = 0.5$ ($\forall i$) and rewrite the privacy constraints in (6) as

$$\frac{a_i(1 - b_j)}{b_i(1 - a_j)} = \frac{1 - b_j}{b_i} \leq e^{r(\epsilon_i, \epsilon_j)} \Leftrightarrow e^{r(\epsilon_i, \epsilon_j)} \cdot b_i + b_j \geq 1$$

Since $a_i = 0.5$, we have $\frac{1-a_i-b_i}{a_i-b_i} = 1$ ($\forall i$), then the total MSE can be represented by

$$\sum_{i=1}^t \frac{nm_i b_i (1 - b_i)}{(a_i - b_i)^2} + \sum_{i=1}^t \sum_{k \in \mathcal{I}_i} c_k^* = \sum_{i=1}^t \frac{nm_i b_i (1 - b_i)}{(0.5 - b_i)^2} + \sum_{k \in \mathcal{I}} c_k^*$$

Therefore, we can obtain the following optimization problem (omit the scaling constant n and the additive constant $\sum_k c_k^*$)

$$\begin{aligned} \min_{0 < b_i < 0.5} \quad f(b) &\triangleq \sum_{i=1}^t \frac{m_i b_i (1 - b_i)}{(0.5 - b_i)^2} \\ \text{s.t.} \quad e^{r(\epsilon_i, \epsilon_j)} \cdot b_i + b_j &\geq 1 \quad (\forall i, j) \end{aligned} \quad (12)$$

Summary of Three Models. opt0 with *non-convex* objective function has $2t$ variables and t^2 *non-linear* privacy constraints. Both opt1 and opt2 have t variables and t^2 *linear* privacy constraints, and the Hessian matrices of their objective functions are positive-definite in the feasible region, thus they are *convex* problems with lower computation complexity. In common cases that only need a small number of privacy levels (i.e., a smaller t), we can use opt0 to obtain the theoretically optimal solution with acceptable computation overhead. But if t is very large, it would be better to use opt1 or opt2 to obtain the near-optimal solution in the shrunk feasible region.

E. Comparison with LDP Mechanisms

In the example discussed in Sec. IV-A, all participants randomly perturb their true answers with a certain probability to protect privacy. Specifically, each participant first generates a vector \mathbf{x} with five bits, where only the position of the truth is 1 and other positions are 0s, then flips each bit with assigned probabilities (depending on the mechanisms) to generate the perturbed vector \mathbf{y} . Finally, the organization aggregates all perturbed vectors from n participants and estimate the counts of these categories by the estimator \hat{c}_i . In Table II, we show

that our proposed mechanism IDUE (solved by opt0) outperforms the state-of-the-art mechanisms (RAPPOR [4] and OUE [7]) under the given privacy levels of inputs, where a smaller total variance $\sum_i \text{Var}[\hat{c}_i]$ indicates a better utility (MSE is equal to the variance for an unbiased estimator). In IDUE, the flipping probabilities for $i = 1$ and $i \neq 1$ are different due to the different privacy levels, while mechanisms satisfying LDP (e.g., RAPPOR and OUE) do not differentiate them. By adjusting the flipping probabilities for different bits, IDUE can achieve the optimal utility with the required protection. The total variance $\sum_i \text{Var}[\hat{c}_i]$ of our mechanism IDUE is in a range because it depends on the distribution of true input data. We can see that the upper bound is still less than that of the existing mechanisms, indicating that our mechanism outperforms others even in the worst-case. For IDUE, the probability of flipping the bit for $i = 1$ may be larger than that in other mechanisms because $\frac{a_1(1-b_j)}{b_1(1-a_j)} = 4 = e^{\epsilon_1}$ ($\forall j$) in RAPPOR and OUE, thus to allow smaller flipping probabilities (i.e., larger $\frac{1-b_j}{1-a_j}$) for $j \neq 1$ under the privacy constraint $\frac{a_1(1-b_j)}{b_1(1-a_j)} \leq e^{\epsilon_1}$ in (6), IDUE needs to increase the flipping probability (hence a larger variance) for $i = 1$ to decrease $\frac{a_1}{b_1}$. This property of IDUE leads to a larger variance for $i = 1$, but smaller flipping probabilities and variance for $i \neq 1$, then the overall utility is improved.

VI. MECHANISM FOR ITEM-SET INPUT

In this section, we consider the item-set input, where the input domain is $\mathcal{D} = \mathcal{P}(\mathcal{I})$, i.e., the power set of \mathcal{I} . If we directly apply the IDUE mechanism developed in Sec. V for this case, each possible set will need to be assigned two perturbation probabilities (for bit-0 and bit-1), therefore the computational cost of solving the optimization problem would be very high because the size of the input domain is 2^m . In this section, we solve the scalability issue by extending the IDUE mechanism with Padding-and-Sampling (PS) protocol to adapt to item-set input. The privacy analysis shows that if mechanism IDUE satisfies MinID-LDP, then the extended one IDUE-PS satisfies MinID-LDP as well. Thus, IDUE-PS has the same computational complexity as IDUE.

A. The Padding-and-Sampling Protocol

Assume the raw data of each user is a set of items, where the number of items in each set can be different. This problem is more challenging than the single-item input even under LDP notion because the user has more than one item, where each item would split privacy budget (reporting all items will lead to large noise in each item and thus bad utility of query).

Algorithm 1 Padding-and-Sampling (PS) [7]

Input: Item-set $x \in \mathcal{D}$ and dummy set $\mathcal{S} = \{m+1, \dots, m+\ell\}$.
Output: One item $x_s \in x \cup \mathcal{S}$

- 1: Set the padded input $x_p \leftarrow x$
- 2: **if** $|x| < \ell$ **then**
- 3: Select $(\ell - |x|)$ dummy items with uniform random from \mathcal{S} and add them into x_p
- 4: **else if** $|x| > \ell$ **then**
- 5: Drop out $(|x| - \ell)$ items with uniform random from x_p
- 6: **end if**
- 7: Sample one item x_s with uniform random from x_p

However, if adopting sampling technique to avoid budget splitting, the different number of items in each user makes the frequency estimation much harder because the sampling probability depends on the number of items of the user which should be kept private. A good solution to address the item-set type of input is the Padding-and-Sampling protocol [7].

Algorithm 1 shows the steps of Padding-and-Sampling protocol, where the item-set $x \in \mathcal{D}$ is padded by a dummy set \mathcal{S} (or truncated) into a new set x_p with a fixed length ℓ and only one item x_s is randomly sampled from the padded set x_p . The fixed length ℓ is a system parameter which will affect the utility in some way (depending on the data distribution). More details of selecting a good ℓ is discussed in [7]. We will discuss how to select ℓ empirically in Sec. VII-B.

B. Mechanism Design and Privacy Analysis

IDUE with Padding-and-Sampling for Item-set Input

By adopting the Padding-and-Sampling (PS) protocol, our previous mechanism IDUE (in Sec. V-B) can be extended for set-valued input. Algorithm 2 shows the steps (sampling, encoding, and perturbing) of our extended mechanism named IDUE-PS, where the data is perturbed according to the sampled item's parameters under the single item case. Fig. 2 shows the diagram of perturbation steps in the user-side and aggregation (on frequency estimation) in the server-side. Since the original itemset input x is padded with some dummy items from a domain \mathcal{S} that is disjoint from the original item domain \mathcal{I} , the item domain is extended to be $\mathcal{I} \cup \mathcal{S}$. We denote the new item domain $\mathcal{I}' = \{1, 2, \dots, m + \ell\}$, where the last ℓ items are dummy items, and the encoded vector \mathbf{x} has $(m + \ell)$ bits. Since each item will be sampled with probability $1/\ell$ from the padded set x_p , the result of frequency estimation needs to be multiplied by the factor ℓ , i.e., $\hat{c}_i = \ell \cdot \frac{c_i - nb_i}{a_i - b_i}$ for $i \in \mathcal{I}$ (we do not need to estimate frequencies of dummy items).

Assume the perturbation probabilities of i -th bit are a_i, b_i , and denote two paramters

$$\alpha_i = \frac{a_i}{b_i}, \quad \beta_i = \frac{1 - a_i}{1 - b_i} \quad (\forall i \in \mathcal{I}') \quad (13)$$

Since $\alpha_i - \beta_i = \frac{a_i - b_i}{b_i(1 - b_i)}$ and $0 < b_i \leq a_i < 1$, we have $1 \leq \beta_i \leq \alpha_i$ ($\alpha_i = \beta_i$ only when $a_i = b_i$). Before proving the privacy guarantee of IDUE-PS, we show the following useful lemma first.

Lemma 2 For any item-set inputs $x, x' \in \mathcal{D}$, and any output y of IDUE-PS (Algorithm 2), the following probability ratio

Algorithm 2 IDUE-PS for Item-Set Input

Input: Item-set $x \in \mathcal{D}$ and dummy set $\mathcal{S} = \{m+1, \dots, m+\ell\}$.
Perturbation probabilities (a_i, b_i) for $i \in \mathcal{I}' = \{1, 2, \dots, m+\ell\}$.
Output: Vector $\mathbf{y} \in \{0, 1\}^{m+\ell}$

- 1: Let $\mathbf{x} = [0, \dots, 0]$ with length $(m + \ell)$
- 2: Sample one item $x_s \in \mathcal{I}'$ by Algorithm 1 and let $\mathbf{x}[x_s] = 1$.
- 3: **for** $k = 1$ to $(m + \ell)$ **do**
- 4: **if** $\mathbf{x}[k] = 1$ **then**
- 5: Randomly draw $\mathbf{y}[k] \sim \text{Bernoulli}(a_k)$
- 6: **else**
- 7: Randomly draw $\mathbf{y}[k] \sim \text{Bernoulli}(b_k)$
- 8: **end if**
- 9: **end for**

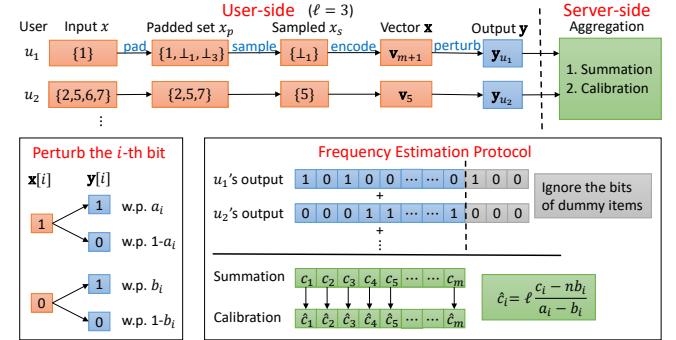


Fig. 2: The IDUE-PS mechanism for item-set input.

is bounded by

$$\frac{\Pr(y|x)}{\Pr(y|x')} \leq \frac{\eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} + (1 - \eta_x) \sum_{i=m+1}^{m+\ell} \frac{\alpha_i}{\ell}}{\eta_{x'} \sum_{j \in x'} \frac{\beta_j}{|x'|} + (1 - \eta_{x'}) \sum_{j=m+1}^{m+\ell} \frac{\beta_j}{\ell}} \quad (14)$$

where $\eta_x = \frac{|x|}{\max\{|x|, \ell\}}$ and $\eta_{x'} = \frac{|x'|}{\max\{|x'|, \ell\}}$

Proof: Denote vector $\mathbf{v}_i = [0, \dots, 0, 1, 0, \dots, 0]$ with length $(m + \ell)$, where only the i -th position is 1 ($i \in \mathcal{I}'$). From the Padding-and-Sampling protocol in Algorithm 1,

$$\begin{aligned} \Pr(y|x) &= \sum_{x_s \in x \cup \mathcal{S}} \Pr(x_s \text{ is sampled}) \cdot \Pr(y|x_s) \\ &= \eta_x \sum_{i \in x} \frac{\Pr(y|x_s=i)}{|x|} + (1 - \eta_x) \sum_{i=1}^{\ell} \frac{\Pr(y|x_s=\perp_i)}{\ell} \\ &= \eta_x \sum_{i \in x} \frac{\Pr(\mathbf{y}|\mathbf{v}_i)}{|x|} + (1 - \eta_x) \sum_{i=m+1}^{m+\ell} \frac{\Pr(\mathbf{y}|\mathbf{v}_i)}{\ell} \end{aligned}$$

where η_x is defined in Lemma 2. On the other hand,

$$\begin{aligned} \Pr(\mathbf{y}|\mathbf{v}_i) &= \Pr(\mathbf{y}[i]|\mathbf{x}[i]=1) \prod_{k \in \mathcal{I}' \setminus i} \Pr(\mathbf{y}[k]|\mathbf{x}[k]=0) \\ &= \frac{\Pr(\mathbf{y}[i]|\mathbf{x}[i]=1)}{\Pr(\mathbf{y}[i]|\mathbf{x}[i]=0)} \Phi = \frac{a_i^{y[i]} (1 - a_i)^{1 - y[i]}}{b_i^{y[i]} (1 - b_i)^{1 - y[i]}} \Phi = \alpha_i^{y[i]} \beta_i^{1 - y[i]} \Phi \end{aligned}$$

where $\Phi \triangleq \prod_{k \in \mathcal{I}'} \Pr(\mathbf{y}[k]|\mathbf{x}[k]=0) > 0$, and α_i, β_i are defined in (13). Since the value of $\mathbf{y}[k]$ is either 1 or 0 and $\alpha_i > \beta_i$, then $\beta_i \leq \frac{\Pr(\mathbf{y}|\mathbf{x}=\mathbf{v}_i)}{\Phi} \leq \alpha_i \quad (\forall i \in \mathcal{I}')$. Thus, we have $\frac{\Pr(y|x)}{\Phi} \leq \eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} + (1 - \eta_x) \sum_{i=m+1}^{m+\ell} \frac{\alpha_i}{\ell}$ and $\frac{\Pr(y|x)}{\Phi} \geq \eta_x \sum_{i \in x} \frac{\beta_i}{|x|} + (1 - \eta_x) \sum_{i=m+1}^{m+\ell} \frac{\beta_i}{\ell}$. Finally,

$$\frac{\Pr(y|x)}{\Pr(y|x')} = \frac{\frac{\Pr(y|x)}{\Phi}}{\frac{\Pr(y|x')}{\Phi}} \leq \frac{\eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} + (1 - \eta_x) \sum_{i=m+1}^{m+\ell} \frac{\alpha_i}{\ell}}{\eta_{x'} \sum_{j \in x'} \frac{\beta_j}{|x'|} + (1 - \eta_{x'}) \sum_{j=m+1}^{m+\ell} \frac{\beta_j}{\ell}}$$

Considering $\frac{\alpha_i}{\beta_j} = \frac{a_i(1-b_j)}{b_i(1-a_j)}$ is the upper bound of $\frac{\Pr(y|x=i)}{\Pr(y|x'=j)}$, the distinguishability of a pair of item-set inputs x and x' in (14) can be regarded as the combined distinguishability of the items that belong to the two sets. The parameter η_x can be explained as the probability of sampling $i \in \mathcal{I}$ from the padded set x_p of input x . If both $|x|$ and $|x'|$ are greater or equal to ℓ (i.e., $\eta_x = \eta_{x'} = 1$), the distinguishability of the pair is averaged only among the items in the set; if not (then $\eta_x < 1$ or $\eta_{x'} < 1$), the distinguishability of the dummy items will be involved since the original set would be padded with dummy items. From Lemma 2, we observe that the distinguishability in IDUE-PS is determined by the privacy levels of the items in the pair of inputs (besides the number of items in the input set), which motivates that IDUE-PS satisfies the notion of MinID-LDP in some way (discussed below).

Privacy Analysis. In (14), the upper bound of the probability ratio $\frac{\Pr(y|x)}{\Pr(y|x')}$ are related to the perturbation probabilities of dummy items, i.e., a_i and b_i for $i = m+1, \dots, m+\ell$. Since the dummy items themselves are not sensitive, we can select some reasonable values as their privacy levels. In this paper, we assume the privacy levels and perturbation probabilities of different dummy items are the same, denoted as $\epsilon_i = \epsilon^*$, $a_i = a^*$, $b_i = b^*$ ($i = m+1, \dots, m+\ell$), then (14) can be rewritten as

$$\frac{\Pr(y|x)}{\Pr(y|x')} \leq \frac{\eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} + (1 - \eta_x)\alpha^*}{\eta_{x'} \sum_{j \in x'} \frac{\beta_j}{|x'|} + (1 - \eta_{x'})\beta^*} \quad (15)$$

where $\alpha^* = \frac{a^*}{b^*}$ and $\beta^* = \frac{1-a^*}{1-b^*}$. We consider the following expression of privacy budget for an item-set

$$\epsilon_x = \ln \left[\eta_x \sum_{i \in x} e^{\epsilon_i} / |x| + (1 - \eta_x)e^{\epsilon^*} \right] \quad (\forall x \in \mathcal{D}) \quad (16)$$

which can be regarded as the combined privacy budget of the items in the set x (the privacy budget of dummy items will be involved when $|x| < \ell$, i.e., $\eta_x < 1$). The combined privacy budget in (16) is larger than the averaged privacy budget $\sum_{i \in x} \epsilon_i / |x|$ because the exponential function $f(\epsilon) = e^\epsilon$ is convex with property $\sum_i k_i f(\epsilon_i) \geq f(\sum_i k_i \epsilon_i)$, where $0 \leq k_i \leq 1$ and $\sum_i k_i = 1$. Based on the results in Lemma 2, we show the fact that IDUE-PS satisfies MinID-LDP.

Theorem 4 *If mechanism IDUE with perturbation probabilities a_i, b_i ($i \in \mathcal{I}$) satisfies MinID-LDP for single-item input with privacy budget $\epsilon_1, \epsilon_2, \dots, \epsilon_m$, i.e.,*

$$\frac{\alpha_i}{\beta_j} = \frac{a_i(1-b_j)}{b_i(1-a_j)} \leq e^{\min\{\epsilon_i, \epsilon_j\}} \quad (\forall i, j \in \mathcal{I}) \quad (17)$$

then IDUE-PS with the same perturbation probabilities will satisfy MinID-LDP for item-set input, i.e.,

$$\frac{\Pr(y|x)}{\Pr(y|x')} \leq e^{\min\{\epsilon_x, \epsilon_{x'}\}} \quad (\forall x, x' \in \mathcal{D}, \forall y) \quad (18)$$

where privacy budget of item-set is defined in (16) and the privacy budget of dummy items $\epsilon^ \in \{\epsilon_1, \epsilon_2, \dots, \epsilon_m\}$,*

Proof: Denote $\alpha_{\max} = \max_{i \in \mathcal{I}} \{\alpha_i\}$, $\beta_{\min} = \min_{j \in \mathcal{I}} \{\beta_j\}$. According to $\alpha_i/\beta_j \leq e^{\min\{\epsilon_i, \epsilon_j\}}$, we have $\alpha^*/\beta_{\min} \leq e^{\epsilon^*}$ and

$\alpha_i/\beta_{\min} \leq e^{\epsilon_i}$ ($\forall i \in \mathcal{I}$). Then (15) can be rewritten as

$$\begin{aligned} \frac{\Pr(y|x)}{\Pr(y|x')} &\leq \frac{\eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} / \beta_{\min} + (1 - \eta_x)\alpha^* / \beta_{\min}}{\eta_{x'} \sum_{j \in x'} \frac{\beta_j}{|x'|} / \beta_{\min} + (1 - \eta_{x'})\beta^* / \beta_{\min}} \\ &\leq \frac{\eta_x \sum_{i \in x} \frac{e^{\epsilon_i}}{|x|} + (1 - \eta_x)e^{\epsilon^*}}{\eta_{x'} \sum_{j \in x'} \frac{1}{|x'|} + (1 - \eta_{x'})} = \frac{\eta_x \sum_{i \in x} \frac{e^{\epsilon_i}}{|x|} + (1 - \eta_x)e^{\epsilon^*}}{1} = e^{\epsilon_x} \end{aligned}$$

where ϵ_x is defined in (16). On the other hand, according to $\alpha_i/\beta_j \leq e^{\min\{\epsilon_i, \epsilon_j\}}$, we have $\beta^*/\alpha_{\max} \geq e^{-\epsilon^*}$ and $\beta_j/\alpha_{\max} \geq e^{-\epsilon_j}$ ($\forall j \in \mathcal{I}$). Then (15) can be rewritten as

$$\begin{aligned} \frac{\Pr(y|x)}{\Pr(y|x')} &\leq \frac{\eta_x \sum_{i \in x} \frac{\alpha_i}{|x|} / \alpha_{\max} + (1 - \eta_x)\alpha^* / \alpha_{\max}}{\eta_{x'} \sum_{j \in x'} \frac{\beta_j}{|x'|} / \alpha_{\max} + (1 - \eta_{x'})\beta^* / \alpha_{\max}} \\ &\leq \frac{\eta_x \sum_{i \in x} \frac{1}{|x|} + (1 - \eta_x) \cdot 1}{\eta_{x'} \sum_{j \in x'} \frac{e^{-\epsilon_j}}{|x'|} + (1 - \eta_{x'})e^{-\epsilon^*}} = \frac{1}{\eta_{x'} \sum_{j \in x'} \frac{e^{-\epsilon_j}}{|x'|} + (1 - \eta_{x'})e^{-\epsilon^*}} \\ &\leq \eta_{x'} \sum_{j \in x'} \frac{e^{-\epsilon_j}}{|x'|} + (1 - \eta_{x'})e^{-\epsilon^*} = e^{\epsilon_{x'}} \end{aligned}$$

The last inequality is obtained by Cauchy-Schwarz inequality. Finally, $\frac{\Pr(y|x)}{\Pr(y|x')} \leq \min\{e^{\epsilon_x}, e^{\epsilon_{x'}}\} = e^{\min\{\epsilon_x, \epsilon_{x'}\}}$. ■

According to Theorem 4, the perturbation probabilities in IDUE-PS for item-set input can be determined with the same way in IDUE, i.e., solving the optimization problems (9) with only $2t$ variables and t^2 constraints to get the optimal solution (t is the number of privacy levels), or the constrained models (11) and (12) with less computational cost to get the near-optimal solution. For the privacy budget of dummy items, theoretically, we can select ϵ^* to be any value from $\{\epsilon_1, \epsilon_2, \dots, \epsilon_m\}$. Though a larger ϵ^* will improve the utility of *dummy* items, the result of frequency estimation for dummy items will be ignored in aggregation because they are not our task. Also, the value of ϵ^* (selected from the original budgets) does not change the optimization problem and the optimal solution because the objective function (only depends on original items) and constraints (only depends on privacy levels) are the same. Therefore, we select $\epsilon^* = \min\{\epsilon_1, \epsilon_2, \dots, \epsilon_m\}$ to guarantee the privacy with smaller budget ϵ_x in (16).

VII. EVALUATION

In this section, we evaluate the performance of frequency estimation of IDUE and compare it with RAPPOR [4] and OUE [6]. Note that RAPPOR and OUE satisfy ϵ -LDP with $\epsilon = \min\{\mathcal{E}\}$, while IDUE and IDUE-PS satisfy \mathcal{E} -MinID-LDP. The perturbation probabilities in IDUE (and IDUE-PS) can be obtained by three optimization models in (9), (11), and (12), denoted by opt0 , opt1 and opt2 respectively.

Applicability of Multiple Privacy Budgets. Though our notion MinID-LDP generally considers different privacy budgets ϵ for different items, in practice, these items can be classified by a small number of categories with distinct privacy levels. For example, thousands of clinical conditions can be classified by three categories including serious diseases, moderate diseases and common symptoms, where only three values of privacy budgets need to be determined according to the applications. We note that the privacy benefit is bounded

TABLE III: Synthetic and Real-world Datasets

Datasets	# Records	# Users (n)	# Items (m)
Power-law	100,000	100,000	100
Uniform	100,000	100,000	1,000
Retail [28]	908,576	88,162	16,470
Kosarak [28]	8,019,015	990,002	41,270
Clothing [29]	192,544	105,508	5,850

by $2 \min\{\mathcal{E}\}$ even when other privacy budgets are higher than $2 \min\{\mathcal{E}\}$ (refer to Lemma 1). In the case of item-set, we consider the privacy budget of a set of items with the form of (16), which is a combination of privacy budgets of items in this set. Theorem 4 shows that the perturbation probabilities of IDUE-PS that satisfies MinID-LDP can be determined by IDUE (where items classified in the same privacy level have the same perturbation probabilities). Therefore, the complexity of our solution, including the number of assigned privacy budgets and computation cost of solving our model, only depends on the number of privacy levels (rather than the domain size of single-item or item-set).

Datasets. We conduct the experiments over two synthetic single-item datasets (with different distributions and domain sizes) and three real item-set datasets (obtained from public data sources), whose parameters are shown in Table III. The data with Power-law distribution is obtained by generating random values from the power-law distribution with the law’s exponent $\alpha = 2$, then scaling and rounding into an integer that belongs to $\mathcal{I} = \{1, 2, \dots, m\}$. The data with Uniform distribution of each user is uniformly generated from $\mathcal{I} = \{1, 2, \dots, m\}$.

Evaluation Metrics. We use the *total* Mean Squared Error (MSE) of all items and the *average* Relative Error (RE) of top k frequent items, defined by

$$\text{MSE} = \sum_{i \in \mathcal{I}} \frac{(\hat{c}_i - c_i^*)^2}{n}, \quad \text{RE}(k) = \frac{1}{k} \sum_{i \in \mathcal{T}(k)} \frac{|\hat{c}_i - c_i^*|}{c_i^*}$$

where \hat{c}_i (resp. c_i^*) is the estimated (resp. true) count of item i , and $\mathcal{T}(k)$ is the set of ground true top k frequent items. We also use the ranking of estimated frequencies to identify top k frequent items and evaluate its *precision* (in Sec. VII-B). All experimental results are averaged with ten repeats.

Setting of Privacy Budget. We consider multiple privacy levels of the inputs, thus we need to assign multiple privacy budgets to them. Assume there are three privacy levels with privacy budget $\{\epsilon, 1.2\epsilon, 2\epsilon\}$ (as default values), where ϵ is the smallest privacy budget. The privacy budget for all items are randomly selected from the three values with a certain budget distribution, where the default distribution is $\{5\%, 5\%, 90\%\}$, and we will change the budget distribution in the experiments to evaluate the impact.

A. Single-item Data

Validation of Theoretical Analysis. Fig. 3 shows the empirical and theoretical results of the MSE of the estimated frequency under Power-law and Uniform distributions. The empirical results (solid lines) are very close to the theoretical results (dashed lines), which validates the correctness of our

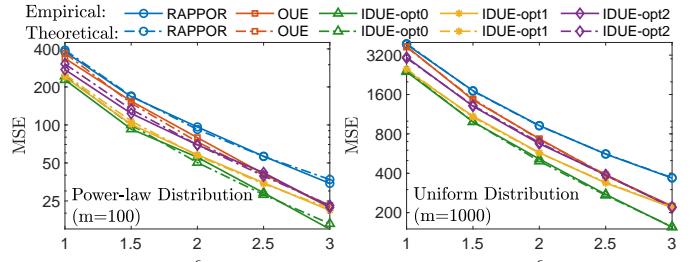


Fig. 3: Comparison of Empirical (dashed lines) and Theoretical (solid lines) results of synthetic data (single-item input).

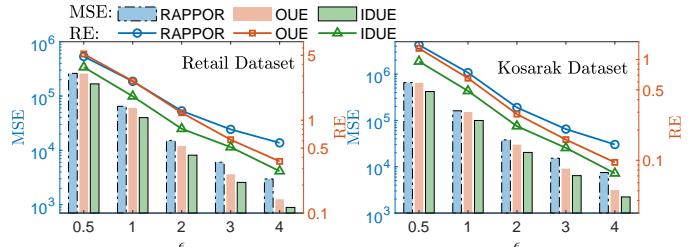


Fig. 4: MSE and RE of real-world datasets (single-item input).

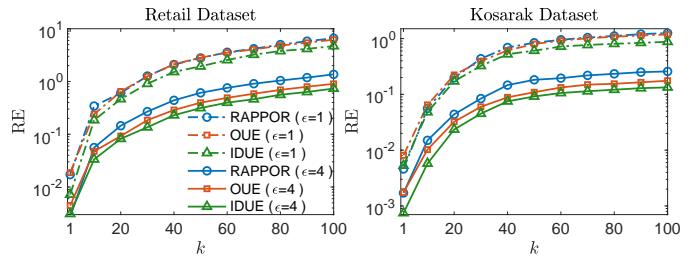


Fig. 5: RE of top k frequent items (varying k).

theoretical analysis. We can observe that mechanisms satisfying LDP and MinID-LDP have relatively similar utility but IDUE with MinID-LDP outperforms RAPPOR and OUE by adjusting the perturbation probabilities for different inputs. For IDUE, the reduced optimization models (i.e., opt1 and opt2) have relatively larger MSEs than the original optimization model (i.e., opt0) due to the further constrained variable space, but they still can provide the near-optimal solution for IDUE with less computational complexity. In the following experiments, we only evaluate IDUE solved by opt0 for simplicity of plots.

Results on Real-world Datasets. Fig. 4 shows the total MSE of all items and average RE of top k frequent items (with $k = 20$) of Retail and Kosarak datasets, where only the first item of each user is considered in the case that each input is a single-item. We also show the results of RE under different k in Fig. 5. The proposed IDUE has the best utility (i.e., smallest MSE and RE of frequency estimation) for all considered ϵ and k .

Influence of Privacy Budget Distributions. The MSE and RE (with $k = 50$) under different budget distributions are shown in Fig. 6, where we only consider two privacy levels (with privacy budgets $\epsilon_1 < \epsilon_2$) and vary the percentage of items whose privacy budget is ϵ_1 (the smaller one). Under a smaller percentage, i.e., only a few of items are more sensitive

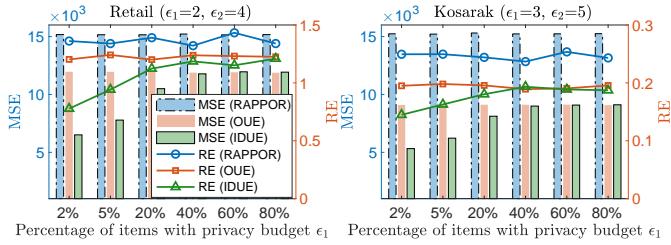


Fig. 6: Under different privacy budget distributions.

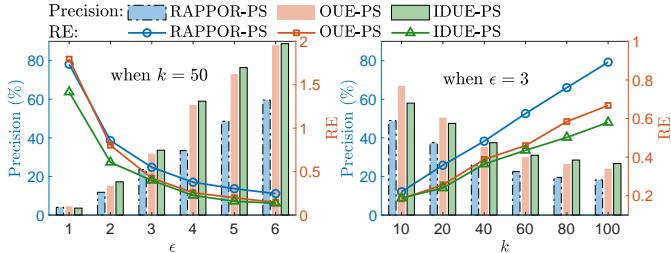


Fig. 7: Precision and RE of Clothing itemset data ($\ell = 2$).

than others, IDUE can get more benefits from our relaxed privacy notion MinID-LDP. However, when the percentage of the more sensitive items is large (such as more than 40%), IDUE has almost the same total MSE and average RE as OUE (which outperforms RAPPOR).

B. Item-set Data

Accuracy of Top Frequent Items Identification. Fig. 7 shows the precision of top k frequent items identification (i.e., the proportion of correct selections over all predicted top frequent items, obtained from the ranking of estimated frequencies) and RE (with the above k) of Clothing dataset under different ϵ and k . We note that each item has privacy budget in $\{\epsilon, 1.2\epsilon, 2\epsilon\}$ with distribution $\{5\%, 5\%, 90\%\}$ (the default one), where ϵ is the budget for a single item, and the privacy budget of each item-set is a combination of items' budgets in the set, defined in (16). The proposed mechanism IDUE-PS has the smallest RE (similar to the previous results) among three mechanisms. But for a smaller ϵ or k (such as $\epsilon = 1$ in the left plot and $k = 10$ or 20 in the right plot), the precision of top frequent items identification may be worse than the precision of OUE-PS (note that for larger ϵ and k , IDUE-PS has the highest precision). Such an observation might be caused by the distinct protection for different items in IDUE-PS, where items with the smallest privacy budget have larger error than the other two mechanisms (which was explained in Sec. V-E). But when ϵ or k is larger, such impact will be mitigated (compared with other mechanisms) because a larger ϵ allows less noise to be added in the perturbation of items with the smallest privacy budget. On the other hand, a larger k generally leads to a lower precision on top frequent items identification because many items in real-world data have the middle ranking, thus the same amount of error on estimated frequencies will make a big difference on the estimated ranking. However, a larger k also makes top items with the smallest privacy budget have more chances to be

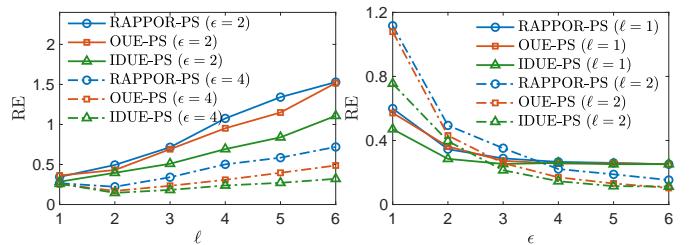


Fig. 8: Varying ℓ and ϵ in Clothing itemset data ($k = 20$).

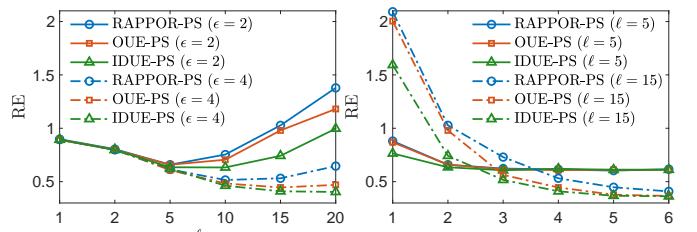


Fig. 9: Varying ℓ and ϵ in Kosarak itemset data ($k = 100$).

selected, thus IDUE-PS can get benefits from the balance of distinct amount of noise of different items (caused by distinct privacy protection levels).

Influence of the Padding Length ℓ for Item-Set Data. The results of Clothing and Kosarak datasets, where each user approximately has 2 and 8 items in average respectively, under different padding length ℓ are shown in Fig. 8 and Fig. 9 (results in Retail data have similar trends as in Kasarak data). We can observe that the optimal or near-optimal ℓ differs in both data distribution (users in Kosarak dataset have more items than in Clothing dataset) and privacy budget ϵ (the optimal ℓ is larger under a larger ϵ). The second observation is caused by the influence of the given ϵ on the trade-off between variance and bias of frequency estimation (i.e., a larger ℓ leads to a larger variance while a smaller ℓ leads to a larger bias [7]). Under a smaller ϵ (i.e., stronger privacy), the error from variance dominates the error from bias, thus a smaller ℓ should be selected to reduce the variance. Similarly, under a larger ϵ , a larger ℓ should be selected to reduce the bias. Also, when fixing a relatively small ℓ (such as $\ell = 5$ in Kosarak), RE does not reduce much with increasing ϵ because ϵ has little influence on the bias (which largely contributes to the error in this case). In [7], ℓ is selected as the 90th percentile of numbers of items of all users (i.e., only depending on data distribution). However, a good ℓ should also depend on ϵ from above discussions. A simple empirical strategy is to select ℓ as the average number of items in each user under a larger ϵ (such as 4), while select ℓ less than the average one under a smaller ϵ (such as 2). The advanced strategy of how to determine the optimal ℓ under a specific ϵ will be our future work.

VIII. DISCUSSIONS

Additional Gain from Incomplete Privacy Policy Graph.

According to Lemma 1, the gain of MinID-LDP compared with LDP is at most twice of the privacy budget, which is caused by the required privacy protection on all pairs of inputs

(i.e., complete graph shown in Fig. 1). However, if some of the pairs do not need to be protected (such incomplete graph can be defined by the secret policy in Blowfish privacy [13]), the gain of MinID-LDP can be larger than $2 \min\{\mathcal{E}\}$ because some inputs might not need to be indistinguishable from the inputs with the smallest privacy budget.

Other Instantiations of ID-LDP. Besides MinID-LDP, other instantiations of ID-LDP can be defined. For example, we can define AvgID-LDP as ID-LDP with the average function, i.e., $r(\epsilon_x, \epsilon_{x'}) = (\epsilon_x + \epsilon_{x'})/2$, which bounds the privacy budget of a pair of inputs by the averaged budget of the two inputs. Similar to MinID-LDP, the notion of AvgID-LDP satisfies sequential composition like Theorem 2. Moreover, the perturbation mechanisms developed in Sec. V and Sec. VI are also applicable to AvgID-LDP.

Benefits of Our Framework. The utility improvement of IDUE is dependent on the utility metrics and the distributions of privacy budget and data. In the case of two different privacy budgets, if items with the smaller budget only have little influence on the utility (generally the number of these items is very small in this case), the utility of IDUE will approach the LDP mechanism with the larger budget. Note that larger noise will be added in the perturbation of the items with the smaller budget to satisfy the privacy constraint, but the impact on utility is very small in this case.

Limitations of Our Framework. First, the amount of benefits of our framework depends on budget distribution. If majority of items have the smallest budget, the benefit obtained from IDUE might be very small (see Fig. 6) because these items greatly affect the utility. Second, the distinct amount of noise for different items may have negative influence on the utility of some applications, such as the precision of top frequent item identification in Fig. 7.

IX. CONCLUSION

In this paper, a new privacy notion named ID-LDP with an instantiation MinID-LDP is proposed to provide input-discriminative protection in the local setting. MinID-LDP is shown to satisfy the sequential composition theorem as LDP and can be regarded as the fine-grained version of LDP. We propose the perturbation mechanism framework IDUE that satisfies ID-LDP, where the perturbation probabilities are solved by the optimization problem with reasonable scale. Then, based on Padding-and-Sampling protocol, the mechanism is extended to apply to item-set input, named IDUE-PS, to solve the scalability and utility problem for the item-set type of input. IDUE-PS is also shown to satisfy MinID-LDP. Finally, experimental results validate the advantage of our privacy notion and mechanisms, compared with the existing ones.

For future work, we will extend our work to handle more complex data types or analysis tasks.

REFERENCES

- [1] C. Dwork, “Differential privacy,” in *ICALP*, 2006, pp. 1–12.
- [2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography Conference (TCC)*, 2006, pp. 265–284.
- [3] R. Chen, H. Li, A. Qin, S. P. Kasiviswanathan, and H. Jin, “Private spatial data aggregation in the local setting,” in *IEEE ICDE*, 2016, pp. 289–300.
- [4] Ú. Erlingsson, V. Pihur, and A. Korolova, “Rappor: Randomized aggregatable privacy-preserving ordinal response,” in *ACM CCS*, 2014, pp. 1054–1067.
- [5] “Learning with privacy at scale,” <https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html>, 2017.
- [6] T. Wang, J. Blocki, N. Li, and S. Jha, “Locally differentially private protocols for frequency estimation,” in *USENIX Security Symposium*, 2017, pp. 729–745.
- [7] T. Wang, N. Li, and S. Jha, “Locally differentially private frequent itemset mining,” in *IEEE S&P*, 2018, pp. 127–143.
- [8] Q. Ye, H. Hu, X. Meng, and H. Zheng, “Privkv: Key-value data collection with local differential privacy,” in *IEEE S&P*, 2019.
- [9] S. Wang, L. Huang, M. Tian, W. Yang, H. Xu, and H. Guo, “Personalized privacy-preserving data aggregation for histogram estimation,” in *IEEE GLOBECOM*, 2015, pp. 1–6.
- [10] M. Andrés, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, “Geo-indistinguishability: Differential privacy for location-based systems,” in *ACM CCS*, 2013, pp. 901–914.
- [11] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, “Broadening the scope of differential privacy using metrics,” in *International Symposium on Privacy Enhancing Technologies Symposium*, 2013, pp. 82–102.
- [12] D. Kifer and A. Machanavajjhala, “A rigorous and customizable framework for privacy,” in *ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, 2012, pp. 77–88.
- [13] X. He, A. Machanavajjhala, and B. Ding, “Blowfish privacy: Tuning privacy-utility trade-offs using policies,” in *ACM SIGMOD*, 2014, pp. 1447–1458.
- [14] M. Bun and T. Steinke, “Concentrated differential privacy: Simplifications, extensions, and lower bounds,” in *Theory of Cryptography Conference*, 2016, pp. 635–658.
- [15] Z. Jorgensen, T. Yu, and G. Cormode, “Conservative or liberal? personalized differential privacy,” in *IEEE ICDE*, 2015, pp. 1023–1034.
- [16] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Local privacy and statistical minimax rates,” in *IEEE FOCS*, 2013, pp. 429–438.
- [17] X. Gu, M. Li, Y. Cheng, L. Xiong, and Y. Cao, “Pckv: Locally differentially private correlated key-value data collection with optimized utility,” in *USENIX Security Symposium*, 2020.
- [18] Y. Nie, W. Yang, L. Huang, X. Xie, Z. Zhao, and S. Wang, “A utility-optimized framework for personalized private histogram estimation,” *IEEE TKDE*, vol. 31, no. 4, pp. 655–669, 2019.
- [19] S. Wang, Y. Nie, P. Wang, H. Xu, W. Yang, and L. Huang, “Local private ordinal data distribution estimation,” in *IEEE INFOCOM*, 2017, pp. 1–9.
- [20] X. Gu, M. Li, Y. Cao, and L. Xiong, “Supporting both range queries and frequency estimation with local differential privacy,” in *IEEE CNS*, 2019, pp. 124–132.
- [21] M. E. Gursoy, A. Tamersoy, S. Truex, W. Wei, and L. Liu, “Secure and utility-aware data collection with condensed local differential privacy,” *arXiv preprint:1905.06361*, 2019.
- [22] T. Murakami and Y. Kawamoto, “Utility-optimized local differential privacy mechanisms for distribution estimation,” in *USENIX Security Symposium*, 2019, pp. 1877–1894.
- [23] F. D. McSherry, “Privacy integrated queries: an extensible platform for privacy-preserving data analysis,” in *ACM SIGMOD*, 2009, pp. 19–30.
- [24] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias,” *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.
- [25] B. Jiang, M. Li, and R. Tandon, “Context-aware data aggregation with localized information privacy,” in *IEEE CNS*, 2018, pp. 1–9.
- [26] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [27] Z. Huang and W. Du, “Oprr: Optimizing randomized response schemes for privacy-preserving data mining,” in *IEEE ICDE*, 2008, pp. 705–714.
- [28] “Kosarak and retail datasets,” <http://fimi.uantwerpen.be/data/>.
- [29] “Clothing dataset,” <https://www.kaggle.com/rmisra/clothing-fit-dataset-for-size-recommendation>.