Hindawi
Security and Communication Networks
Volume 2020, Article ID 8829523, 29 pages
https://doi.org/10.1155/2020/8829523

Review Article

A Comprehensive Survey on Local Differential Privacy

Xingxing Xiong,[1] Shubo Liu,[1] Dan Li,[1,2] Zhaohui Cai,[1] and Xiaoguang Niu[1]

[1]*School of Computer Science, Wuhan University, Wuhan 430072, China*
[2]*Hubei Water Resources Research Institution, Wuhan 430070, China*

Correspondence should be addressed to Shubo Liu; liu.shubo@whu.edu.cn

Received 7 June 2020; Revised 7 August 2020; Accepted 12 August 2020; Published 8 October 2020

Academic Editor: Angel M. Del Rey

With the advent of the era of big data, privacy issues have been becoming a hot topic in public. Local differential privacy (LDP) is a state-of-the-art privacy preservation technique that allows to perform big data analysis (e.g., statistical estimation, statistical learning, and data mining) while guaranteeing each individual participant's privacy. In this paper, we present a comprehensive survey of LDP. We first give an overview on the fundamental knowledge of LDP and its frameworks. We then introduce the mainstream privatization mechanisms and methods in detail from the perspective of frequency oracle and give insights into recent studied on private basic statistical estimation (e.g., frequency estimation and mean estimation) and complex statistical estimation (e.g., multivariate distribution estimation and private estimation over complex data) under LDP. Furthermore, we present current research circumstances on LDP including the private statistical learning/inferencing, private statistical data analysis, privacy amplification techniques for LDP, and some application fields under LDP. Finally, we identify future research directions and open challenges for LDP. This survey can serve as a good reference source for the research of LDP to deal with various privacy-related scenarios to be encountered in practice.

1. Introduction

With the development of Information Technology, people have been enjoying more convenient life and higher quality services. This is especially prominent in the era of big data. However, privacy issues have been becoming a hot topic in public in recent years and may become a major obstacle to the long-term development of information technology and will influence the public's acceptance of technologies. To avoid this curse, privacy-preserving individual data have become a top priority for governments and organizations in the world. The European Union is a pioneer in the context of individual data privacy preservation. The EU passed the General Data Protection Regulation (GDPR) (https://gdpr-info.eu/) in 2016. Hereafter, many countries outside of the EU have successively enacted and activated laws or regulations on privacy protection of individual data, e.g., Brazil's General Data Protection Law (PDPL) (https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf) and India's Personal Data Protection Bill (PDPB) (https://www.prsindia.org/sites/default/files/bill_files/Personal%20Data%20Protection%20Bill%2C%202019.pdf). In addition, the Cyber Security Law of the People's Republic of China (CSL) (https://en.wikipedia.org/wiki/China_Internet_Security_Law) has been in effect since 2017.

However, it is insufficient for privacy preservation to only rely on these laws and regulations and it requires the support of privacy protection techniques. Recently, differential privacy [1] is proposed and regarded as a prestandard privacy preservation technique for quantifying privacy. The technique can provide strong and provable privacy preservation, which is reflected in three aspects: providing stringent and tunable privacy preservation level since adding or deleting any record makes no difference in results; defending against powerful attack model since it does not rely on knowing how much background knowledge the attacker has; and possessing complete and provable mathematical theoretical foundations. Compared to the anonymous-based privacy preservation methods (e.g., k-anonymity [2], l-diversity [3], and t-closeness [4]) that require assumptions on some specific attack and background knowledge, differential privacy has been becoming a hotpot in academic and industry fields.

Traditional differential privacy (called centralized differential privacy, CDP) [1] has focused on the private statistical data publication from the collected raw data. In particular, a centralized data server collects the raw data from data providers and publishes the private statistics information perturbed by differential privacy mechanisms. In the centralized setting, there is an assumption that the centralized data server (data collector) must be trusted and secure so that it will not steal or disclose participants' sensitive information. However, the setting is often confronted with many privacy leak problems in practical deployment applications. For example, in 2018, American social network giant Facebook's user data were gotten access by the British analytics firm Cambridge Analytica, which resulted in at least 80 million pieces of user information being leaked (https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal). The world's biggest hotel chain Marriott International's approximately 500 million customers' data were hacked from its guest reservation database (https://www.wired.com/story/marriott-hack-protect-yourself). Security magazine listed 2019's Top 12 Data Breaches (https://www.securitymagazine.com/articles/91366-the-top-12-data-breaches-of-2019) which included America's largest real estate title insurance company First American Financial Corp exposing about 885 million individuals' transaction records and Chinese Job seekers MongoDB data Breach bringing about disclosing more than 200 million job candidates' information records. Such information leakage incidents have emerged one after another in recent years so that people are increasingly worried about the security of their personal information.

Recently, local differential privacy (LDP) [5, 6] has been proposed which is a more rigorous differential privacy technique for personal information privacy preservation. LDP is based on the assumption that the data collector is untrusted. Specifically, in the setting, each participant locally perturbs her/his raw data with a differential privacy mechanism and transfers the perturbed version to the untrusted server (data collector/aggregator). Until all participants' perturbed data are received, the server calculates the statistics and publishes the statistical result. Benefiting from its advantage, LDP was widely researched and applied by the industrial field as soon as it was proposed. There are several state-of-the-art practical deployment applications of LDP for privately collecting fashionable statistics so that LDP is used by hundreds of millions of people every day. RAPPOR [7] was developed by Google, which is based on the Randomized Response and Bloom Filters to implement privately collecting Chrome's setting of massive users. This is the first and pure client-based practical privacy solution, which no longer needs the participation of a trusted third-party and takes control of user's data in their own hands. Apple [8] announced that LDP was deployed in the iOS system, which is documented in patent application and research paper. The technique is based on the LDP with other techniques including using the Fourier transform technique to filter out information and sketching techniques to reduce the dimensionality of domain. Telemetry data collection under LDP was deployed in Windows 10 by Microsoft, which makes use of histograms to collect data about systems and applications over time [9]. Besides, Samsung [10] has also investigated the LDP technique, but it is not clear that it has been deployed in practice.

Up to now, a large number of studies on LDP have been emerging, but there is little literature focused on the review of local differential privacy. In 2017, Ye et al. [11] conducted a survey of LDP that merely focused on the frequency estimation, mean value estimation, and the design of perturbation model. In 2018, Cormode et al. [12] presented a tutorial for simply reviewing current research that addressed some partly related problems within the LDP model. In 2019, Bebensee et al. [13] also presented an overview of different LDP algorithms for problems such as locally private heavy hitter identification and spatial data collection. Li and Ye [14] performed a small survey on LDP in a seminar talk that introduced its fundamental behind these practical deployment systems, reviewed its current research landscape, and recognized some open challenges in LDP. In addition, a recent review work [15] focused on the survey on LDP for securing Internet of vehicles, and it is different from the above-mentioned review studies since it is a survey of LDP-based application. Nevertheless, these studies of survey literature did not conduct a comprehensive and detailed survey on local differential privacy and, there does not exist a comprehensive and systematic survey on the existing research studies of LDP.

To this end, we do this work on a comprehensive survey of the latest research process and directions of LDP. In particular, we first describe the fundamental knowledge on LDP including its definition, fundamental mechanisms, properties, metrics methods, and comparisons between LDP and CDP, as well as introduce the frameworks of LDP. We then introduce the mainstream privatization mechanisms (methods) for basic statistical estimations (e.g., frequency estimation and mean estimation) and review the complex statistical estimations (e.g., distribution estimation over multivariate data, set-valued data, and graph structure data). Moreover, we conduct an extensive literature survey of current research circumstances of LDP which focuses on the private statistical learning/inferencing, statistical data analysis under LDP, privacy amplification techniques for LDP, and application fields under LDP. Finally, we discuss several open challenges for LDP. In summary, the main contributions of this work are as follows:

(i) We first give an overview on the fundamental knowledge and frameworks of LDP and introduce mainstream privacy mechanisms for frequency oracles protocols.

(ii) We then conduct an extensive literature review of basic statistical distribution estimation and complex statistical distribution estimation problems under LDP.

(iii) We further give a comprehensive review of current research circumstances about LDP in terms of statistical learning/inferencing, statistical data analysis, privacy amplification techniques, and application fields.

(iv) We identify several open challenges in local differential privacy on the basis of our review.

1.1. Organization. The remainder of this paper is organized as follows. We introduce an overview of the fundamental knowledge of local differential privacy in Section 2. The frameworks of local differential privacy preservation are presented in Section 3. Mainstream privatization mechanisms for frequency oracle protocols are explored in Section 4. In Section 5, we introduce in detail the existing researches on private statistical estimation under LDP. Section 6 is about the current research circumstances for LDP. Ultimately, some open challenges in the field are identified in Section 7.

2. Fundamental Knowledge

2.1. Definition of LDP. We formally introduce the LDP model and give a brief overview of the related concept of LDP. The LDP model fully considers the possibility of an untrusted server or data collector (aggregator) stealing or comprising the privacy of participants (users). In the setting, there are a large number of participants. The i-th participant holds a sensitive value (raw data) v_i in domain D. These participants interact with the data collector (aggregator) so that the collector obtains statistical information about the distribution of data while guaranteeing the individual's privacy. More specifically, a participant locally perturbs his private value v_i utilizing a perturbed algorithm and sends its output $A(v)$ to the collector. Then the collector reconstructs the collected reports to acquire an effective statistical analysis result. The formal privacy requirement is that algorithm $A(\bullet)$ satisfies the following property.

Definition 1 (Local Differential Privacy). A randomized algorithm is (ε, δ)-local differential privacy ((ε, δ)-LDP), where $\varepsilon \geq 0$ and $0 \leq \delta \leq 1$, if and only if any pair of input values $v, v' \in D$, for all $O \subseteq \text{Range}(A)$,

$$\Pr[A(v) \in O] \leq e^{\varepsilon} \cdot \Pr[A(v') \in O] + \delta, \tag{1}$$

where $\text{Range}(A)$ denotes the set of all possible outputs of the algorithm A. If $\delta = 0$, the algorithm A satisfies pure (strict) local differential privacy (pure LDP), namely, ε-LDP. If $\delta > 0$, the algorithm satisfies approximate (relaxed) local differential privacy (approximate LDP), namely, (ε, δ)-LDP.

Similar to the centralized setting, LDP controls plausible deniability for any two values so that the algorithm A satisfies ε-LDP or (ε, δ)-LDP. In brief, given the output of the privacy algorithm, it is almost impossible to infer which value the input data is. In the centralized setting, the privacy of algorithm A is defined by the concept of a neighbor dataset, it requires a trusted data collector to collect data and privately release analysis results of the data. In the Local setting, each participant can independently deal with her/his data; to be specific, the privacy process is transferred from the data collector (center) to individual participant (local). Therefore, LDP essentially avoids privacy attacks from the untrusted data collector.

Some studies have presented several related definitions of LDP. Jiang et al. [16] proposed a novel definition of Localized Information Privacy (LIP) for frequency estimation of context-aware data, which relaxes the classic LDP by introducing a context-aware knowledge (priors) to increase statistics utility. LIP is identical to ε-adversarial privacy (ε-AP) [17] and implies ε-Mutual Information Privacy (ε-MIP) [18]. It imposes a constraint on the ratio between the prior and posterior. The result demonstrates that the LIP mechanism performs better tradeoffs between privacy and utility than the classic LDP. However, LIP has no enough significant advantage for uniform prior distribution and has only been applied to privacy-preserving frequency estimation for binary alphabet at present. Recently, Acharya and Kairouz et al. [19] proposed a general definition of context-aware LDP for the practical applications where not all elements of data domain are equally sensitive.

2.2. Fundamental Mechanisms for LDP

2.2.1. Randomized Response Mechanism. Similar to the perturbation mechanism for CDP, namely, Laplace mechanism [20] and Exponential mechanism [21], Randomized Response (RR) mechanism is the primary perturbation mechanism for LDP. In 1965, Warner [22] first proposed randomized response as a research method for privacy survey interview about embarrassing or illegal behaviors. Its main idea is based on the plausible deniability responding to sensitive information (such as criminal behavior or sexuality) to maintain individuals' privacy. Formally, each individual participant holds her/his own data that may be one element (a database of size 1) to answer a binary question in a differentially private manner. To better specify, reporting a binary data (single bit) by the RR, each participant reports the true value with probability p and the nontrue value with probability $1 - p$. The method satisfies $(\ln(p/1 - p))$-LDP. The basic randomized response is called W-RR.

To explain the RR with an example, we assume a specific scenario: the National Bureau of Statistics wants to learn how many AIDS patients live in China without clearly knowing who is an AIDS patient. The Bureau claims each participant to answer the question, "Are you an AIDS patient?" in the way to flip a biased coin. If it comes up heads, response truthfully. If it comes up tails, response "Yes" with probability and "No" with probability $1 - p$. Finally, the National Bureau of Statistics cannot distinguish which participants are the AIDS patients but still can estimate the proportion of the AIDS patients with high confidence. The RR is viewed as a well-designed noise-adding mechanism to obfuscate individual data while enabling the computation of aggregate statistics.

2.2.2. Laplace Mechanism. Laplace mechanism [20] sometimes is also adopted in the LDP. Herein, we will briefly introduce it. It is first proposed and adopted for CDP. Since it is suitable to perturb the numerical data. In particular, given any function $f\colon \mathbb{D} \longrightarrow \mathbb{R}^d$, the Laplace mechanism is defined as

$$M(f(x), \varepsilon) = f(x) + (Z_1, \ldots, Z_d), \tag{2}$$

where Z_i are independent and identically distributed (i.i.d) random variables drawn from the Laplace distribution with scale $\Delta f/\varepsilon$, namely $\mathrm{Lap}(\Delta f/\varepsilon)$, and Δf is the ℓ_1-sensitivity of the function f and can capture the magnitude by a single value data can infect the function f in the worst case. Δf is defined as

$$\Delta f = \max_{x,y\in\mathbb{D}} \|f(x) - f(y)\|_1. \tag{3}$$

2.2.3. Other Mechanisms. The above-mentioned RR mechanism is the most fundamental mechanism for LDP. Besides, there are information-theory-based perturbation mechanisms that include information compression and distortion mechanisms.

Sarwate et al. [23] first studied the tradeoff of privacy-utility under differential privacy from the distortion-measured perspective and proposed an information-distortion-based LDP mechanism (aka. Distortion). Suppose the participant passes the m-ary alphabets data $X = \{x\}^k$ through a private channel $Q(z\mid x)$ to produce $Z = \{z\}^k$. In order to guarantee utility, the distortion parameter δ is set for constraining the deviation of X and Z.

$$\max_{P\in \mathrm{P}} \mathbb{E}_{p\times Q}[d(X,Z)] \leq \delta. \tag{4}$$

The deviation is measured by Hamming distortion, and the distortion is defined as:

$$d(X,Z) = \frac{1}{k}\sum_{i=1}^{k} d(x_i, z_i). \tag{5}$$

The private channel $Q(z\mid x)$ is determined by the distortion parameter δ, where δ is constant. The channel $Q(z\mid x)$ is defined as

$$Q(z\mid x) = \begin{cases} 1-\delta, & z = x, \\ \delta/(m-1), & z \neq x, \end{cases} \tag{6}$$

where m is denoted alphabets size, that is, the number of possible values of. The private channel satisfies ε-LDP and $\varepsilon = \log(m-1) + \log((1-\delta)/\delta)$.

From the same team with Sarwate, Kalantari et al. [24] further studied the tradeoff of privacy-utility under LDP and Hamming distortion over an arbitrary set of finite-alphabet source distributions. Meanwhile, they analyzed and demonstrated the optimal differential privacy mechanisms for three class source distributions classified by different assumptions on prior knowledge, respectively.

Xiong et al. [25] proposed a novel information compression-based perturbation mechanism for LDP (aka Compression) and studied the fundamental tradeoffs between privacy, compression, and utility. Each participant possesses a length k sequence of data $X = (x_1, \ldots, x_k)$, where $x_i \in \mathbb{X}$, and $\mathbb{X}$ is an input discrete alphabet; then pass X through channel Q to release perturbed version $Z = (z_1, \ldots, z_k)$, where $z_i \in \mathbb{Z}$, and $\mathbb{Z}$ is an output discrete alphabet, $|\mathbb{Z}| \leq |\mathbb{X}|$. The privatized and compressed process satisfied ε-locally differential private, and the compression ratio ρ is defined as

$$\rho = \frac{\log_2|\mathbb{Z}|}{\log_2|\mathbb{X}|}. \tag{7}$$

In order to limit the information distortion caused by compression, the method also introduced a target constrain δ to constrain the expected hamming distortion, that is,

$$\max_{P\in\mathbb{P}} \mathbb{E}_{P\times Q}[d(X,Z)] \leq \delta, \tag{8}$$

where $\mathbb{E}_{P\times Q}[d(X,Z)] = \sum_{z_i,z_i} P(x_i)Q(z_i\mid x_i)d(x_i,z_i)$ is the expected hamming distortion, X and Z is independent and identically distributed (i.i.d) distribution drawn from P and Q, respectively.

Given a distribution set, compression ratio ρ, and distortion constrain δ, finding the channel Q which yields the optimal $\varepsilon^*(\mathbb{P}, \rho, \delta)$ has been turned into a convex optimization problem and can be solved via the bisection method.

Table 1 shows the advantages and disadvantages of these perturbation mechanisms under LDP. Randomized response mechanism has become the main perturbation mechanism under LDP due to its high scalability. It is mainly used to measure the relationship between input and output data. Laplace mechanism is a noise-adding perturbation mechanism under LDP and suitable for the perturbation of numerical data. In the distortion-based perturbation mechanism, privacy budget ε depends on the size of attribute candidates, and the two have positive proportional relationship. As a result, when the size of attribute candidates is large, the degree of privacy protection will go down. The compression-based perturbation mechanism is merely applicable to low-dimensional data due to the high dimension of which is determined by the Cartesian product of input and output alphabet size. The feasible set of Q is exponentially proportional to the number of dimensions. As a result, the higher the dimension, the higher the corresponding computing complexity. The last two mechanisms mainly consider the relationship between input and output data from the perspective of information loss. Note that we mainly survey the randomized response mechanism in this paper.

2.3. Properties of LDP. Local differential privacy possesses three properties: sequential composition property, parallel composition property, and postprocessing property. These properties are defined as follows.

Property 1 (sequential composition). Suppose n mechanisms $\{M_1, \ldots, M_n\}$ satisfy ε_i-LDP, respectively, and are sequentially computed on the private data, then a mechanism combined by $(M_1, \ldots, M_n)$ in some order satisfies $(\sum_{i=1}^{n} \varepsilon_i)$-LDP.

Property 2 (parallel composition). Suppose n mechanisms $\{M_1, \ldots, M_n\}$ satisfy ε_i-LDP, respectively, and are computed on a disjoint subset of the private data, then a mechanism formed by $(M_1(D_1), \ldots, M_n(D_n))$ satisfies $(\max(\varepsilon_i))$-LDP.

Property 3 (postprocessing property). If mechanism M_1 satisfies ε-LDP, for any mechanism M_2, even may not satisfies LDP, the composition of M_1 and M_2, namely, $M_2(M_1(\cdot))$ satisfies ε-LDP.

Table 1: Comparisons of perturbation mechanisms under LDP.

Mechanism	Advantages	Disadvantages
Randomized response mechanism [22]	High scalability; low communication (only statistical computation); and computation cost (encoding processing)	Only categorical data
Laplace mechanism [20]	High scalability and low computation cost (only computation of noise addition)	Only numerical data and sensitivity problem
Distortion-based mechanism [23]	Information loss quantification	Only low-dimensional data and additional computation cost of information-distortion
Compression-based mechanism [25]	Information loss quantification	High computation and communication complexity, additional computation cost of compression rate, and distortion on information

2.4. Metrics Method. LDP protocols make the untrusted aggregator obtain the statistical estimation from a large number of participants' sensitive data while guaranteeing the privacy of the individual participants. Its main goal is to pursue the tradeoff between the utility (accuracy) on the aggregator side and the privacy on the participant side. In [26], the work formalized the differences between aggregate and individual information by the knowledge of information theory and presented several novel information-theoretic metrics for utility and privacy in the LDP, such as worst-case privacy, limit behavior for utility, and privacy-utility tradeoff. In this paper, we only focus on the survey study on LDP privacy (the systematic survey of privacy metrics including differential privacy is presented in [27]); herein, we believe present common utility metrics under LDP to analyze the method (protocol) of LDP. We classify them into the error-based metrics [28] and the information-theoretic metrics [29].

2.4.1. Error-Based Metrics for Utility. In statistical estimation (e.g., frequency and distribution estimation), the error-based metrics are common metrics for utility. From the perspective of adversary, the metrics describe the error between the private observation Z and the original (real) observation X. The error-based metric is described as

$$\text{Error} = \mathbb{E}\left[\|X - Z\|_p\right], \tag{9}$$

where is called the ℓ_p-norm. In general, one chooses the ℓ_1- and ℓ_2-norms as the metric of utility (error), when ℓ_1-norm is set, the error is called the Mean Absolute Error (MAE) or L_1 error; when ℓ_2-norm is set, the error is the Mean Squared Error (MSE) or L_2 error. The MAE and MSE are defined, respectively, as

$$\begin{aligned} \text{MAE} &= \frac{1}{|X|} \sum_{x\in X, z\in Z} |(z - x)|, \\ \text{MSE} &= \frac{1}{|X|} \sum_{x\in X, z\in Z} (z - x)^2. \end{aligned} \tag{10}$$

2.4.2. Information-Theoretical Metrics. For the general scenario (purposes), information loss metrics is adopted to quantify the error between the original data and private data. For a specific purpose such as data mining, machine learning, or statistical analysis task. The private data are used as the input to these tasks. The quality of the task result is evaluated by the accuracy or error rate compared to the original data. Herein, we mainly introduce several common information-theoretical metrics in terms of utility for the general scenarios, such as statistical estimation.

(1) Mutual information: Mutual information is treated as a general measurement of statistical data utilities. The mechanism aims to maximize mutual information. The mutual information between X and Z is defined as

$$I(X, Z) = \sum_{x,z} p(x, z) \log \frac{p(x, z)}{p(x)p(z)}. \tag{11}$$

(2) KL divergence: The *Kullback–Leibler* divergence (KL divergence) is commonly adopted to measure the similarity between distributions. The KL divergence between X and Z is

$$D_{KL}(X\|Z) = \sum_{x} p(x) \log\left(\frac{p(x)}{p(z)}\right), \tag{12}$$

where X and Z denote the original distribution and private distribution.

2.5. Comparison of LDP and CDP

2.5.1. Privacy Model. In the CDP model, there is an assumption that the data collector is trustworthy; each participant sends private data to the collector. Once a data analyst requests a query, the data collector responds to the request using an algorithm that satisfies certain level differential privacy. In the LDP setting, the assumption that the data collector is untrusted is more practical. To preserve the data privacy, the differential privacy algorithm is transferred from the data collector to each participant. Each participant solely perturbs her/his data in light of the differential privacy algorithm and then sends the perturbed data to the collector. Similarly, when a data analyst launches a query, the data collector responds to the request over the collected data. The framework of local and centralized differential privacy model is shown in Figure 1 (participant is abbreviated as particip).

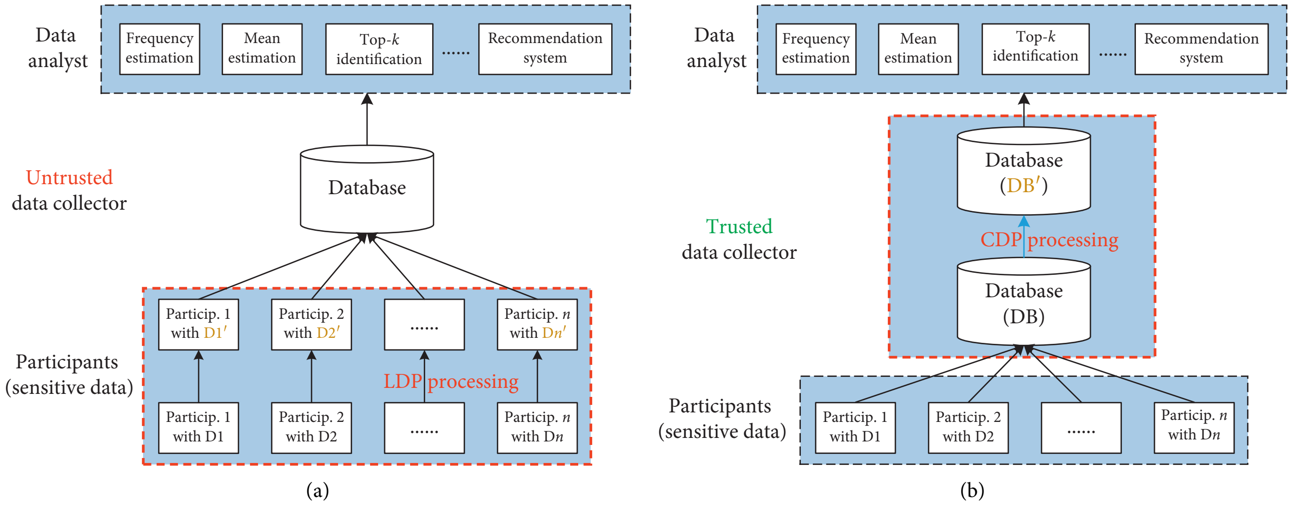


Figure 1: Framework of local and centralized differential privacy model: (a) LDP model; (b) CDP model.

2.5.2. Perturbation Mechanism. Since different privacy is a probabilistic model, any differentially private perturbation mechanism must be random. In the CDP setting, Laplace mechanism and Exponential mechanism are the two most popular perturbation mechanisms. Generally, the former is applicable in numerical data by adding controlled noise to the query function. But the latter is used to perturb categorical data. It is worth noting that the two perturbation mechanisms are closely related to the definition of the global sensitivity of the query function. The global sensitivity is defined on the neighbor datasets with at most one record difference, which makes it impossible for an attacker to infer individual record from the statistical results, that is, hiding individual record in the results. In the LDP setting, each participant perturbs his data by himself and then sends the randomized data to the data collector, and any two participants learn nothing about each other's private data. Since each participant only holds her/his own local data, there is no global sensitivity, but local sensitivity in the LDP. Hence, the key to adopting the above two mechanisms for achieving LDP is to determine the local sensitivity. So far, the randomized response mechanism is the mainstream LDP mechanism, one reason is that it does not depend on the concept of sensitivity.

2.5.3. Privacy-Utility Tradeoff. Compared with the CDP model, it is more challenging to achieve a reasonable privacy-utility tradeoff under the LDP model. There are two main reasons. On one hand, LDP requires the addition of higher noise than what is required by CDP; that is, the former requires a lower bound of noise magnitude $\Omega(\sqrt{n}/\varepsilon)$, where n is the number of participants. However, the latter requires $O(1/\varepsilon)$. On the other hand, LDP has no assumption of neighborhood constraint on participants' data as inputs, once when the data domain is very large, LDP brings about a significant reduction of utility.

3. Local Differential Privacy Preservation Framework

Identical to the CDP, the LDP has two frameworks of privacy preservation, interactive and noninteractive framework [5, 30, 31]. To formalize the two frameworks, let $X_1, \ldots, X_n \in \mathbb{X}$ be original data and $Z_1, \ldots, Z_n \in \mathbb{Z}$ be the corresponding private (perturbed) data. The original random variables $\{X_i\}_{i=1}^n$ and private observations $\{Z_i\}_{i=1}^n$ are linked by a privacy preservation mechanism channel Q. We refer to Q as a pipeline from the original to the privatized data. The relationship is formalized by the conditional probability. The framework structure of LDP is shown in Figure 2.

3.1. Noninteractive Framework. The noninteractive framework [30] is that Z_i depends only on X_i and not on any other private variables Z_j for $j \neq i$. The noninteractive conditional independence structure is

$$X_i \longrightarrow Z_i \text{ and } Z_i \perp \left\{X_j, Z_j, j \neq i\right\} \middle| X_i, \tag{13}$$

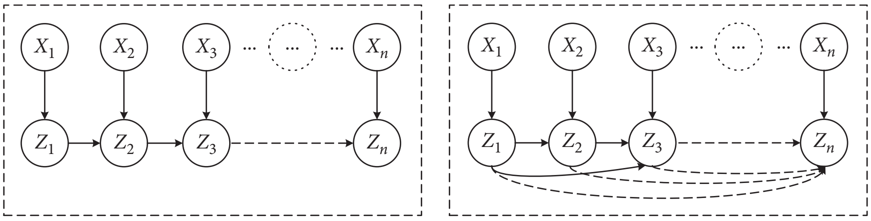
where ⊥ denotes the symbol of independent relation.

We have

$$\sup_{S\in\sigma(Z)} \sup_{x,x\prime\in\mathbb{X}} \frac{Q\left(Z_i \in S \middle| X_i = x\right)}{Q\left(Z_i \in S \middle| X_i = x'\right)} \leq e^{\varepsilon}. \tag{14}$$

There are two noninteractive protocols including shared/public randomness protocol and local/private randomness protocol. For the former, the protocol requires the generation of shared randomness on the server. For the latter, the protocol has no shared randomness in the noninteractive framework.

3.2. Interactive Framework. There are two interactive settings in this framework, namely, sequentially interactive setting and fully interactive setting [32, 33]. In sequentially



(a)

(b)

FIGURE 2: Structure diagram of LDP Framework. (a) Noninteractive framework; (b) one-round interactive framework.

interactive setting [30], participants release actually one message each in sequence, and these messages may rely on the previously released messages. The number of interactive rounds of the setting is substantially limited by the number of participants but can be fewer. Note that each participant outputs a message at most once. In fully interactive setting [34], each participant may release any number of messages with arbitrary dependencies on the other previously known messages, and there is no restriction on the number of rounds.

3.2.1. Sequentially Interactive Setting. The privatized variable Z_i depends on both the corresponding original variable X_i and the former $i-1$ private variables $Z_{<i} = (Z_1, \ldots, Z_{i-1})$, while it is independent on the $X_{<i} = (X_1, \ldots, X_{i-1})$. We assume it is one-round sequentially interactive, the interactive conditional independence structure is

$$\{X_i, Z_{<i}\} \longrightarrow Z_i \text{ and } Z_i \perp X_j \Big| \{X_i, Z_{<i}\}, \quad \text{for } j \neq i, \tag{15}$$

where $\perp$ denotes a symbol of independent relation. For a given privacy parameter $\varepsilon \geq 0$, $Z_{<i}$ is a ε-LDP observation of $X_{<i}$ if for all $z_{<i} = (z_1, \ldots, z_{i-1})$ and $x, x' \in X$, we have

$$\sup_{S \in \sigma(Z)} \frac{Q_i\left(Z_i \in S \,\middle|\, X_i = x, Z_{<i} = z_{<i}\right)}{Q_i\left(Z_i \in S \,\middle|\, X_i = x\prime, Z_{<i} = z_{<i}\right)} \leq e^{\varepsilon}, \tag{16}$$

where $\sigma(Z)$ denotes a possible σ-field on Z.

3.2.2. Fully Interactive Setting. The privatized variable Z_i depends on both the corresponding original variable $X_{\leq n} = (X_1, \ldots, X_n)$. We assume it is one-round fully interactive, the interactive conditional independence structure is

$$\{X_{\leq n}\} \longrightarrow Z_i \text{ and } Z_i \perp X_j \Big| \{X_{\leq n}\}, \quad \text{for } j \neq i, \tag{17}$$

where $\perp$ denotes a symbol of independent relation. For a given privacy parameter $\varepsilon \geq 0$, pair of samples $x_{\leq n}, x'_{\leq n} \in \mathbb{X}^n$ differing in at most a single element, we have

$$\sup_{S \in \sigma(Z)} \frac{Q_i\left(Z_i \in S \,\middle|\, X_{\leq n} = x_{\leq n}\right)}{Q_i\left(Z_i \in S | X_{\leq n} = x'_{\leq n}\right)} \leq e^{\varepsilon}, \tag{18}$$

where $\sigma(Z)$ denotes a possible σ-field on Z.

In summary, the definitions of the above three frameworks capture a property of plausible deniability: whatever the private data Z has released, it is nearly equally as likely to have derived from one variable as with any other. However, the most critical difference between interactive and noninteractive framework is the correlation over the perturbed data. The former is applied to a scenario (situation) where the current private value has a dependency on the previous $i-1$ private value, such as health data analysis. The latter is applied to a scenario where the private value is independent of any previous perturbed value, such as shopping data analysis. The difference between the sequentially and fully interactive frameworks resides in the number of interactions for each participant. The former requires one-time interaction, while the latter does not have this limitation that each participant can interact any number of times. Therefore, the fully interactive framework satisfies the need of the practical scenarios.

4. Mainstream Privatization Mechanisms for Frequency Oracles Protocols

We first introduce several mainstream privatization mechanisms or methods for Frequency Oracles (FO) protocols that conduct the frequency estimation of any value in the domain, as FO is the basic block for frequency estimation that is the most basic problem of statistical estimation. Since these mechanisms are the foundation for achieving LDP, they can also extend to other problems besides frequency estimation. In general, the LDP protocol consists of three steps: Encoding, Perturbing, and Aggregating, where the Encoding and Perturbing steps are in the side of participant, and the Aggregating steps are in the side of server or analyst. Note that the Aggregating step of FO is to estimate the frequencies.

4.1. Generalized Randomized Response Mechanism. The Randomized response is a fundamental mechanism in LDP. In 1965, Water et al. [22] first proposed the basic randomized response called W-RR for the case of binary alphabets, also termed as 1 bit RR (one-bit RR) or binary RR. Recently, Kairouz et al. [29, 35] proposed a staircase mechanism, termed as K-RR, which is designed for multivariate alphabets case. Wang et al. [36] generalized them into the generalized randomized response (GRR). GRR is formalized as follows: k is denoted as the candidate size from the domain. The perturbation function of GRR is defined as

$$\Pr\left[\text{Perturb}_{\text{GRR}}(v) = z)\right] = \begin{cases} p = \dfrac{e^{\varepsilon}}{e^{\varepsilon} + k - 1}, & \text{if } z = v, \\ q = \dfrac{1}{e^{\varepsilon} + k - 1}, & \text{if } z \neq v. \end{cases} \tag{19}$$

GRR satisfies ε-LDP since the ratio of p and q is equal to e^{ε} or $1/e^{\varepsilon}$.

The GRR-based protocol is simple due to no encoding in the encoding step, it takes directly the original value as input into the Perturbing step, namely, GRR perturbation mechanism.

4.2. RAPPOR and Its Variants

4.2.1. RAPPOR. The Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) [7] is designed to collect aggregate statistics from data providers with strong LDP. It builds on the idea of memoization and performs two rounds of randomized response (one of them is memoization step) to guarantee one-time and longitudinal privacy of each participant. Assume that one holds a value. The specific process in the LR is as follows:

(1) Encoding with Bloom filtering: A report value v is encoded into a fixed-length Bloom filter B_0 represented as a bit vector of length h. B_0 is defined as follows:

$$B_0[i] = \begin{cases} 1, & \text{if } \exists H \in \mathbb{H}, \text{ s.t.}, H(v) = i, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

where $\mathbb{H} = \{H_1, H_2, \ldots, H_m\}$ is denoted as a set of m hash functions.

(2) Permanent RR (PRR): Permanent RR is the first round of RR. To defend inference attack on the participant's real answer by reporting the value multiple times, PRR is adopted to transfer B_0 into "noisy" answer B_1 which is then memoized and reused to replace the real answer every time, where the bit vector B_1 is referred to as the permanent randomized response (PRR) for v. Each bit in the PRR $B_1[i]$ is determined:

$$B_1[i] = \begin{cases} 1, & \text{with probability} \dfrac{1}{2} f, \\ 0, & \text{with probability} \dfrac{1}{2} f, \\ B_0[i], & \text{with probability} 1 - f, \end{cases} \quad (21)$$

where f is a user-tunable longitudinal privacy guarantee parameter.

(3) Instantaneous RR (IRR): The collector requests a report each time, each bit $B_1[i]$ in the PRR is delivered into the next round of randomized response to compute the output response bit vector by IRR. Bit vector B_2 is initialized and set to 0; then each bit $B_1[i]$ is set with probabilities.

$$\Pr(B_2[i] = 1) = \begin{cases} p, & \text{if } B_1[i] = 1, \\ q, & \text{if } B_1[i] = 0, \end{cases} \quad (22)$$

where the privacy parameters of IRR are p and q. The smaller the gap between p and q, the stronger the privacy guarantee. Note that B_0, B_1, B_2 have the same fixed-length.

(4) Reporting: Send the generated vector B_2 to the server.

Lastly, the server receives all reports and computes the aggregation estimation. Since the use of hashing of Bloom Filter brings the problem of potential collisions that two different values are hashed to the same position of the Bloom Filter. RAPPOR introduces the notion cohort to group the participants into several cohorts where each cohort uses an unequal set of hash functions so that the collision is limited to within the scope of one cohort. However, potential collisions can still happen and impact estimation accuracy. These complexities make the aggregation algorithm more complicated; RAPPOR adopts LASSO and linear regression to estimate the frequency distribution.

4.2.2. Typical Variants of RAPPOR. Basic RAPPOR [7] is a variant of RAPPOR in which original value can be directly mapped to one bit in a bit (binary) vector instead of exploiting bloom filter in RAPPOR. It is suitable for the case of a relatively small and well-defined set of values (e.g., relatively small deterministic set of strings). *One-time RAPPOR* [7] is a modification without the process of Instantaneous RR. Because a value is reported once, it no longer needs to withstand inference attacks on multiple reports, which is regarded as a longitudinal attack. Note that defending the longitudinal attack by PRR in Basic RAPPOR and RAPPOR assume that participant's value does not change over time. *Basic One-time RAPPOR* [7] (also called *K-RAPPOR* [35]) is the combination of the two variations. Optimized Unary Encoding (OUE) [36] is an optimized Basic One-time RAPPOR method by choosing the optimal parameters $p = 1/2$ and $q = 1/(e^{\varepsilon} + 1)$ to guarantee low variance.

O-RAPPOR is proposed to solve the problem of unknown domain. Building on the same idea of O-RR that adopts the hashing cohort, O-RAPPOR firstly groups participants into different cohorts with an independent hash function. In any cohort, then it samples a hash $H \in \mathbb{H}$ without replacement to map the values from participants into Bloom Filter (BF). Lastly, it adopts the perturbation process of RAPPOR to perturb the BLOOM. The comparison of RAPPOR and its variants is shown in Table 2.

4.3. Matrix-Based Randomized Response Mechanisms

4.3.1. Random Matrix Projection (SHist). Bassily and Smith [37] proposed a Succinct histogram (abbreviated as SHist) protocol based on the Random Matrix Projection technique. The randomized matrix is generated by the Johnson–Lindenstrauss (JL) lemma which indicates that any point set $\mathbb{V}$ in high-dimensional space (k) can be randomly projected into the lower-dimensional Euclidean space $(O(\log(k)))$ so that the distortion of pairwise distance can

Table 2: Comparison of RAPPOR and its variants.

Method	Encoding	Longitudinal reports	Perturbing
Basic RAPPOR [7]	Unary (bit vector)	Multiple times	PRR + IRR
RAPPOR [7]	Hash unary (bloom filter)	Multiple times	PRR + IRR
One-time RAPPOR [7]	Hash unary (bloom filter)	Once	PRR
Basic one-time RAPPOR [7] (K-RAPPOR) [35]	Unary (bit vector)	Once	PRR
OUE [36]	Unary (bit vector)	Once	PRR (optimal parameters)
O-RAPPOR [35]	Unknown domain (bloom filter)	Multiple times	PRR + IRR

be controlled by the confidence parameter δ. The Johnson–Lindenstrauss lemma is formalized as follows.

Theorem 1 (Johnson–Lindenstrauss lemma). *Given $0 < \delta < 1$ and a set of points V in $\mathbb{R}^d$, there exists a linear map $\Phi\colon \mathbb{R}^d \longrightarrow \mathbb{R}^m$ for $m = O(\ln(|V|)/\delta^2)$ and all $\mathbf{u}, \mathbf{v} \in V$ such that*

$$(1 - \delta)\|\mathbf{u} - \mathbf{v}\|_2^2 \leq \|\Phi\mathbf{u} - \Phi\mathbf{v}\|_2^2 \leq (1 + \delta)\|\mathbf{u} - \mathbf{v}\|_2^2. \qquad (23)$$

The parameter m is the dimension of the new space. In fact, it could be any value that is large enough to make the lemma work. However, in SHist, m is decided by the error bound defined as the maximal Hamming distance between the estimation and real frequency of any domain. Note that m can be set to $m = \ln(k + 1)\ln(2/\beta)/\delta^2$, where $\delta = \sqrt{\ln(2k/\beta)/(\varepsilon^2 n)}$ and n is denoted the number of participants.

Firstly, the server generates a public randomized matrix $\Phi \in \{-1/\sqrt{m}, 1/\sqrt{m}\}^{m\times k}$ uniformly at random and shares the matrix Φ to each participant. Then each participant encodes his/her real value by $\text{Encode}(v) = \langle s, x\rangle$, where s is chosen uniformly at random from $[m]$, and x is the v-th element of the s th row of Φ, namely, $x = \Phi[s, v]$ that is a binary value. Next, the participant perturbs the encoding value x by a one-bit RR (WRR) mechanism, the Perturbation step is $\text{Perturb}(\langle s, x\rangle) = \langle s, a\rangle$, where

$$a = \begin{cases} 1 \cdot c_\varepsilon \cdot m \cdot x, & \text{with probability}\, p = \dfrac{e^\varepsilon}{(e^\varepsilon + 1)}, \\[2ex] -1 \cdot c_\varepsilon \cdot m \cdot x, & \text{with probability}\, q = \dfrac{1}{(e^\varepsilon + 1)}, \end{cases} \qquad (24)$$

$$\text{where}\, c_\varepsilon = \frac{e^\varepsilon + 1}{e^\varepsilon - 1}.$$

Lastly, the server receives all reports $\langle s_j, a_j\rangle$, the estimation for $i \in [k]$ is computed by $\widetilde{c}(i) = \sum_j a_i \cdot \Phi[s_j, i]$.

The Random Matrix Projection (SHist) mechanism is proposed to solve the heavy communication cost problem similar to the RAPPOR method. In the protocol, the communication cost of each user is $O(1)$ and the computation cost is $O(k)$ for calculating one row of the Matrix. The computation cost of the server includes $\Theta(km)$ for generating the Matrix and $\Theta(mk)$ for calculating the estimations.

4.3.2. Hadamard Randomized Response (HRR). Hadamard Randomized Response [10, 38, 39] (HRR) is a special Matrix-based RR protocol. HRR is built on the Hadamard transform (Discrete Fourier transformation) matrix described by an orthogonal, symmetric matrix Φ of dimension $k \times k$ (where k is a power of 2), each entry in Φ is defined by

$$\Phi[i][j] = \frac{1}{\sqrt{k}}(-1)^{(i,j)}, \qquad (25)$$

where (i, j) is the bitwise dot product of the binary representations of the numbers i and j.

The encoding, perturbing, and aggregation steps in [10, 38] are the same as the Randomized Matrix Projection (SHist). The HRR is optimal to the SHist, but its limitation is that HRR is only suitable to the case that k is the power of 2. However, the HRR in [39] (called HR mechanism) is a novel Hadamard matrix-based RR without the shared matrix (randomness) on the server, which is a more general of GRR; thus, its implement is similar with the GRR.

4.4. Local Hash Mechanism

4.4.1. Binary Local Hashing (BLH). Similar to RAPPOR, the key idea underlying Binary Local Hashing (BLH) is adopting a hashing technique to lower communication cost. More specifically, Hash the original value into a smaller domain or candidate of size $k < |D|$ and then apply the UE method to the hashed value. Differencing from the RAPPOR, BLH method eliminates the potential effect of collisions by grouping users themselves into different cohorts. BLH method is logically equivalent to the *Randomized Projection Matrix*-based method proposed by Bassily and Smith [37]. Given a general (universal) family of Hash functions $\mathbb{H}$, each hash function $H \in$ can map an input value $V \in [\mathrm{d}]$ to one bit $v = H(v)$ into a bit. The general (universal) property of the family is formalized by

$$\forall x, y \in D, x \neq y\colon \Pr_{H\in\mathbb{H}}[H(x) = H(y)] \leq \frac{1}{2}. \qquad (26)$$

Firstly, an original value v can be encoded as Encode $(v) = (H, b)$, where $H \in \mathbb{H}$ is randomly selected in a uniform way and $b = H(v)$. Then the perturbed value b' is computed by

$$\Pr(b' = 1) = \begin{cases} p = \dfrac{e^\varepsilon}{e^\varepsilon + 1}, & \text{if}\, b = 1, \\[2ex] q = \dfrac{1}{e^\varepsilon + 1}, & \text{if}\, b = 0. \end{cases} \qquad (27)$$

Lastly, the support function of BLH is $\text{Support}_{\text{BLH}}((H, b)) = \{v \mid H(v) = b\}$. Note that each encoded value (H, b) supports half of the original values that are hashed by H to b since the reported information is only a single bit.

4.4.2. Optimal Local Hashing (OLH). Building on the observation that information loss is concentrated in the encoding step for BLH method since the reported value is merely one bit. Wang et al. [36] proposed the generalized improvement of BLH, named Optimal Local Hashing (OLH) that alternately hashes each encoded value into a value in $[g] (g \geq 2)$. The selection of g is essential since larger g indicates that more information is being maintained in the encoding step but more information loss in the perturbing (RR) step. In addition, the optimal parameter $g = e^{\varepsilon} + 1$ is analytically given by them.

In OLH, given a universal hash function family $\mathbb{H}$, each hash function $H \in \mathbb{H}$ can map any original value to a value in $[g]$. Firstly, the original value v is encoded as $\text{Encode}(v) = (H, x)$, where $H \in \mathbb{H}$ is randomly selected in a uniform way and $x = H(v)$. Then the perturbed value x' is computed by

$$\Pr(x' = v) = \begin{cases} p = \dfrac{e^{\varepsilon}}{e^{\varepsilon} + g - 1}, & \text{if } x = v, \\ \\ q = \dfrac{1}{e^{\varepsilon} + g - 1}, & \text{if } x \neq v. \end{cases} \tag{28}$$

Lastly, the support function of OLH is $\text{Support}_{\text{OLH}}((H, x)) = \{v \mid H(v) = x\}$. Noting that OLH is based on GRR and its estimation variance is independent of domain size $|D|$, so it is suitable for the large domain.

In addition, Wang et al. [40] further proposed Symmetric Local Hashing (SLH) to theoretically improve the privacy-utility tradeoff in shuffling-based privacy amplification method, and the slight difference between SLH and OLH is that the optimal parameter is set to be $e^{\varepsilon/2} + 1$ rather than $e^{\varepsilon} + 1$ in OLH.

4.5. Other Mechanisms for Frequency Oracle

4.5.1. O-RR Mechanism. Kairouz et al. [35] proposed the O-RR mechanism for the case of the unknown domain. The O-RR mechanism is an improvement of GRR mechanism. The protocol is based on the idea of hashing and cohorts adopted in RAPPOR. The use of hashing cannot know the domain in advance rather than only consider the hashed domain (the number of selected hash functions), and the adoption of cohorts can further reduce the probability of collision of hashing. Intuitively, each participant i is assigned to a cohort c_i selected uniformly from $C = \{1, \ldots, C\}$. The participants in use the same hash function of a cohort to divide original domain V into k disjoint subsets. For any original value v_i, its encoding value is calculated by

$$x_i = \text{HASH}_{c_i}(v_i) \bmod k = \text{HASH}_{c_i}^{(k)}(v_i). \tag{29}$$

Note the hash functions of between different cohorts are mutually independent so as to reduce the probability of collusion. In the extreme case of the same string in different cohorts, the probability of collision is approximate $1/k$. Formally, $\Pr(x_i = x_j \mid c_i \neq c_j, v_i = v_j) \approx 1/k$.

Next, ORR uses the GRR mechanism for the perturbation process. For any value v_i, the perturbed value r_i is determined by

$$\Pr(r_i \mid v_i) = \frac{1}{C(e^{\varepsilon} + k - 1)} \cdot \begin{cases} e^{\varepsilon}, & \text{if } \text{HASH}_{c_i}^{(k)}(v_i) = r_i, \\ \\ 1, & \text{if } \text{HASH}_{c_i}^{(k)}(v_i) \neq r_i. \end{cases} \tag{30}$$

4.5.2. O-RAPPOR Mechanism. Kairouz et al. [35] also proposed the O-RAPPOR mechansim to solve the problem of unknown domain. Building on the same idea of O-RR that adopts the hashing cohort, O-RAPPOR firstly groups participants into different cohorts where each cohort has an independent h-hash Bloom filter. Formally, for any value v_i, the encoded value x_i is determined by

$$x_i[j] = 1, \text{and } j = \text{HASH}_{c_i}(v_i) \bmod k = \text{HASH}_{c_i}^{(k)}(v_i), \tag{31}$$

where $\text{HASH}_{c,h}^{(k)}$ are a group of hC mutually independent hash functions and $h' \in [1, \ldots, h]$. Note the encoded value x_i is a vector. The perturbation process of O-RAPPOR is adopting the basic one-time RAPPOR mechanism (k-RAPPOR).

4.5.3. k-Subset Mechanism. Previous protocols only consider the case where the perturbation output is a single value over the original domain when any input value is perturbed. However, k-Subset mechanism [41–43] is proposed for the case where the perturbation output is a set of values over the original domain. For any input value $v \in D$ (where $d = |D|$) and output value set $S \subseteq D$ with size k, the perturbation process of k-subset mechanism is determined by

$$\Pr(S \mid v) = \frac{d}{(ke^{\varepsilon} + d - k)C_d^k} \begin{cases} e^{\varepsilon}, & \text{if } v \in S, \\ \\ 1, & \text{if } v \notin S. \end{cases} \tag{32}$$

According to the symmetric property of the above conditional probabilities in the mechanism, the k-subset mechanism is implemented by the reservoir sampling which is exploited to design the mechanism rather than direct sampling from the output domain with size C_d^k. The key of this method is randomly selecting k or $k{-}1$ elements from $D/\{v\}$ so that the computational and storage overheads are both $O(d)$. k-Subset mechanism is regarded as a general form of randomized response mechanism. For example, 1-Subset mechanism is equivalent to the GRR mechanism.

4.6. Classification of Privatization Mechanisms for Frequency Oracle. We classified these protocols by the different encoding methods proposed by Wang et al. [36] into direct encoding method, unary encoding method, and local hashing method. We show the classification of the above-mentioned privatization mechanisms or method for frequency oracle protocols in Table 3. Note that the

Table 3: The classification of privatization mechanisms.

Encoding method	Methods (protocols or mechanisms)
Direct encoding	GRR [36] (*k*-RR [35]), *k*-subset [41, 42]
Unary encoding	Baisc RAPPOR [7], one-time RAPPOR [7], *k*-RAPPOR [35], RAPPOR [7], O-RAPPOR [35], O-RR [35], OUE [36]
Local hashing	RMP(SHist) [37], OLH [36], SLH [40], HRR [10, 38], HR [39]

classification of these privatization mechanisms will be conducted in Section 5.1.1.

5. Locally Differentially Private Statistical Estimation

The majority of existing studies focus on applying LDP to complex data and/or analysis tasks including basic statistical estimation (e.g., frequency estimation and mean estimation) and complex statistical estimation (e.g., joint distribution and distribution estimation over complex data.). In the following sections, we describe those studies in detail.

5.1. Basic Statistical Distribution Estimation under LDP. In this section, we focus on two basic statistical estimation problems under LDP: frequency estimation for categorical attribute data and mean estimation for numerical attribute data. **Frequency estimation** (discrete distribution estimation) is one of the most popular tasks in the LDP model. In the problem, the participant's sensitive data are represented as categorical numerical data (discrete data), and the server aims to estimate the frequency distribution on the population. In particular, for any $v \in [k]$, the server wants to estimate frequency of $v \in [k]$: $f(v) = |\{i\colon x_i = v\}|/n$, namely, $\widehat{f}(v)$. **Mean estimation** is also a popular statistical task. Likewise, the server wants to estimate the mean of $\overrightarrow{x} = \langle x_i \rangle_{i\in[n]}$, namely, $u(\overrightarrow{x}) = (1/n)\sum_{i\in[n]} x_i$, and its mean estimation is denoted as $\widehat{u}$.

5.1.1. Frequency Estimation under LDP. In the previous , we have introduced several fundamental Frequency Oracle mechanisms (protocols) that are proposed for frequency estimation with single categorical attribute data. Herein, we will analyze and summarize the advantages (Pros.) and disadvantages (Cons.) of these algorithms. Meanwhile, we further revisit several optimization mechanisms under LDP for frequency estimation. In Table 4, we show the comparisons of the above-mentioned FO mechanisms for frequency estimation with single categorical attribute data. Note that there are few studies of frequency estimation for multiple categorical attributes data [10, 44]. They adopted the sampling technique for utility improvement by allocating the privacy budget to the sampling attribute or attributes set. The two methods are simple and elegant solutions to handle multiple categorical attributes.

In addition, Jia et al. [45] proposed the Calibrate method to calibrate item frequencies generated from an existing LDP algorithm by incorporating the prior knowledge about noise in estimated item frequencies and true item frequencies through statistical inference, and the method can effectively reduce estimation errors. Previous studies on LDP for frequency estimation have focused on an assumption that all personal data are equally sensitive. However, it brings excessive obfuscation into utility loss. Based on this analysis, Murakami et al. [46] introduced the notion of Utility-optimized LDP (ULDP) and studied the two different settings. They further proposed utility-optimized RR and RAPPOR mechanisms providing ULDP for the setting where all users employ the same obfuscation mechanism and proposed a personalized ULDP mechanism with semantic tags for the anther setting where the difference between sensitive and nonsensitive data may vary among users.

5.1.2. Mean Estimation under LDP. We review the existing studies on the problem of estimating the mean over the numerical attribute data under ε-LDP. The problem setting: each participant holds a vector t_i with d numerical attributes. The server (aggregator) aims to estimate the mean of each attribute over all n participants. For simplicity, suppose that each numerical attribute's value situates in the range $[-1, 1]$. Currently, the mainstream solutions include *Laplace-Mechanism*-based noise-adding methods and *Randomized Response-Mechanism*-based randomization methods.

A naive solution is applying **Laplace Mechanism** [20] to each attribute value in each participant's vector. In particular, $t_i^*[j] = t_i[j] + \mathrm{Lap}(2d/\varepsilon)$, $j \in [d]$, and $i \in [n]$, where $\mathrm{Lap}(\lambda)$ denotes a random noise drawn from Laplace distribution with scale λ, with the probability density function $\mathrm{pdf}(x) = (1/2\lambda)\exp(-|x|/\lambda)$. Once the aggregator receives all perturbed tuples, it obtains the average $(1/n)\sum_{i=1}^{n} t_i^*[j]$ as the mean estimate of attribute j. Obviously, the estimate is unbiased, as the Laplace noise $\mathrm{Lap}(2d/\varepsilon)$ in each attribute has zero mean. The method incurred $O(d/\varepsilon\sqrt{n})$ expected error which is proportional to the number of attributes d and could be overly large if there are a large number of attributes.

Soria-Comas and Domingo-Ferrer [47] proposed an optimal variant of Laplace mechanism (called as **SCDF LM**) for multidimensional data that achieve improved accuracy result. Afterward, Geng and Kairouz et al. [48] proposed **Staircase mechanism** that is a geometric mixture of uniform random variables. It is used to replace the Laplace mechanism for performance improvement. SCDF's Laplace Mechanism and the Staircase mechanism both are the special cases of piece-wise constant probability density distribution, respectively:

$$\mathrm{pdf}(r_i = x) = \begin{cases} C_0 e^{-i\varepsilon}, & \text{if} x \in [-m - 2(i+1), -m - 2i], i \in \mathbb{N}, \\ C_0, & \text{if} x \in [-m, m], \\ C_0 e^{-i\varepsilon}, & if x \in [m + 2i, m + 2(i+1)], i \in \mathbb{N}. \end{cases} \tag{33}$$

Table 4: Comparisons of frequency oracle mechanisms for frequency estimation under LDP.

Method	Encode	Randomness	Asymptotic bound error	Candidate	Communication cost	Computation cost	Pros and cons
k-RR [35] GRR [36]	Direct	Local	$O(k\sqrt{k}/\varepsilon\sqrt{n})$	Known	P: Θ $(\log(k))$ S: Θ	P: O (1) S: O $(n+k)$	Pros: no encoding, predigest the process; lower candidate size can achieve higher utility; cons: low utility in low privacy regime
O-RR [35]	Unary (bloom filter)	Local	$O(k\sqrt{k}/\varepsilon\sqrt{n})$	Unknown	P: O (h) S: O (nh)	P: O (k) S: Linear regression	Pros: open candidate; cons: low utility in low privacy regime, high computation cost due to regression
RAPPOR [7]	Unary (bloom filter)	Local	$O(k/\varepsilon\sqrt{n})$	Known	P: O (h) S: O (nh)	P: O (k) S: LASSO and linear regression	Pros: lower error, lower storage cost, support big candidate; cons: consider Bloom filter parameter settings, high computation cost due to regression
k-RAPPOR (basic one-time) [7]	Unary	Local	$O(k/\varepsilon\sqrt{n})$	Known	P: Θ (k) S: O (nk)	P: $O(k)$ S: $O(n+k+(nk/e^{\varepsilon/2}))$	Pros: lower error, lower storage overhead, simpler and faster implement; cons: consider parameter settings of Bloom filter
OUE [36]	Unary	Local	$O(k/\varepsilon\sqrt{n})$	Known	P: Θ (k) S: O (nk)	P: O S: $O(n+k+(nk/e^{\varepsilon}))$	Pros: lower error, lower storage cost, lower computation cost and easier to implement; cons: larger candidate lead to higher communication cost
O-RAPPOR [35]	Unary (bloom filter)	Local	$O(k/\varepsilon\sqrt{n})$	Unknown	P: Θ (h) S: O (nh)	P: O (k) S: linear regression	Pros: open candidate, higher utility, lower storage overhead; cons: need consider parameter settings of bloom filter
k-Subset [41, 42]	Direct	Local	$O(k/\varepsilon\sqrt{n})$	Known	P: Θ (k) S: O (nk)	P: O (k) S: $O(n+k+(nk/e^{\varepsilon}))$	Pros: better sample complexity and higher utility; cons: higher communication and computation cost due to set output
RMP (SHist) [37]	Binary	Public (shared matrix)	$O(\sqrt{\log k}/\varepsilon\sqrt{n})$	Known	P: O (1) S: O (n)	P: O S: O (nk)	Pros: lower communication cost; cons: unstable query accuracy due to the noise from RMP matric
HRR [10, 38]	Binary	Public (shared matrix)	$O(\sqrt{\log k}/\varepsilon\sqrt{n})$	Known	P: O (1) S: O (n)	P: O (k) S: O	Pros: lower communication cost; cons: unable query accuracy due to the noise from RMP matric
BLH [36]	Binary	Local and public	$O(\sqrt{\log k}/\varepsilon\sqrt{n})$	Known	P: O (1) S: Θ $(\log(n))$	P: O S: O (nk)	Same with the RMP method

Table 4: Continued.

Method	Encode	Randomness	Asymptotic bound error	Candidate	Communication cost	Computation cost	Pros and cons
OLH [36]	Binary	Local and public	$O(\sqrt{\log k}/\varepsilon\sqrt{n})$	Unknown	P: O (1) S: Θ (log(n))	P: O (k) S: O (nk)	Pros: higher utility in the setting big candidate size, lower communication cost; cons: unstable accuracy due to the noise from RMP matric
HR [39]	Binary	Local	$O(k/\varepsilon\sqrt{n})$	Known	P: O (log (k)) S: (O ($n$log (k))	P: O (k) S: O ($n+k$)	Pros: obtain efficient computation complexity due to Fast Walsh–Hadamard transform; cons: unstable accuracy due to the noise from encoding

If $C_0 = \varepsilon/4$ and $m = 2(1 - e^{-\varepsilon} - \varepsilon e^{-\varepsilon})/\varepsilon(1 - e^{-\varepsilon})$, the distribution is the SCDF's Laplace distribution, if $C_0 = (1 - e^{-\varepsilon})/(2m + 4e^{-\varepsilon} - 2me^{-\varepsilon})$ and $m = 2/(1 + e^{\varepsilon/2})$, the distribution is the Staircase distribution. The two methods are achieved by adding the noise from the two piece-wise constant distributions to each real attribute value rather than noise from Laplace distribution. Their asymptotic errors are approximate $O(2de^{\varepsilon/2}/(e^{\varepsilon} - 1)\sqrt{n})$. It is worth noting that the optimality result in [48] does apply to the case with unbounded inputs.

Duchi et al. [30] first proposed a method of mean estimation for numerical attributes under LDP by using randomized response mechanism, hereafter referred to as **MeanEst**. The main idea is that each participant's exact vector (tuple) $t_i \in [-1, 1]^d$ is mapped into a perturbed vector $t_i^* \in \{-B, B\}^d$, where B is a constant determined by ε and d. Its calculation process is rather complicated and is shown by

$$B = \begin{cases} \dfrac{2^d + 2^{d-1}\cdot(e^{\varepsilon} - 1)}{C_{d-1}^{(d-1)/2}\cdot(e^{\varepsilon} - 1)}, & \text{if } d \text{ is odd}, \\[2ex] \dfrac{2^d + \left(2^{d-1} - C_d^{d/2}/2\right)\cdot(e^{\varepsilon} - 1)}{C_{d-1}^{d/2}\cdot(e^{\varepsilon} - 1)}, & \text{otherwise}. \end{cases} \tag{34}$$

MeanEst first computes and generates a random vector $v \in \{-1, 1\}^d$ by choosing each element $v[j]$ independently from the distribution

$$\Pr[v[j] = a] = \begin{cases} \dfrac{1 + t_i[j]}{2}, & \text{if } a = 1, \\[2ex] \dfrac{1 - t_i[j]}{2}, & \text{if } a = -1. \end{cases} \tag{35}$$

Then it returns a perturbed vector $t_i^* \in \{-B, B\}^d$ with the probability $e^{\varepsilon}/(1 + e^{\varepsilon})$ so that $t_i^* \cdot v \geq 0$, or returns it with the probability $1/(1 + e^{\varepsilon})$ so that $t_i^* \cdot v < 0$. The perturbation mechanism is the 1 bit randomized response. Lastly, the aggregator receives all participants' perturbed vector t_i^* and estimates the mean statistics for each attribute. The MeanEst achieves $O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$ asymptotic error bound.

Building on the MeanEst method, Nguyen and Xiao et al. [10] proposed the **Harmony-mean** method by adopting the Sampling technique. Harmony-mean is more efficient than MeanEst for the multidimensional numerical data, namely, $d \geq 2$, since each participant only transmits 1 bit to the server in Harmony-mean, instead of d-bit in MeanEst.

Given an exact vector $t_i \in [-1, 1]^d$, Harmony-mean returns a perturbed vector $t_i^* \in \{-(e^{\varepsilon} + 1)/(e^{\varepsilon} - 1), 0, (e^{\varepsilon} + 1)/(e^{\varepsilon} - 1)\}^d$ where only one bit (attribute $j \in [d]$) is non-zero. In particular, it first initializes the perturbed vector $t_i^* = [0]^d$, and then chooses j uniformly at random from $[d]$, and $t_i^*[j]$ is sampled from the following distribution:

$$\Pr[t_i^*[j] = a] = \begin{cases} \dfrac{t_i[j]\cdot(e^{\varepsilon} - 1) + e^{\varepsilon} - 1}{2\cdot(e^{\varepsilon} + 1)}, & \text{if } a = \dfrac{e^{\varepsilon} + 1}{e^{\varepsilon} - 1}\cdot d, \\[2ex] \dfrac{-t_i[j]\cdot(e^{\varepsilon} - 1) + e^{\varepsilon} - 1}{2\cdot(e^{\varepsilon} + 1)}, & \text{if } a = -\dfrac{e^{\varepsilon} + 1}{e^{\varepsilon} - 1}\cdot d. \end{cases} \tag{36}$$

The perturbed vector t_i^* is a binary vector, and only its j-th bit is nonzero value, so the server only receives 1 bit information from each participant indicating its sign and rescales the bit by parameters ε and d. Therefore, the communication cost of Harmony-mean is one d-th of that of B. The former is the same as the latter in terms of asymptotic error, namely, $O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$.

To further improve the performance, Wang and Xiao et al. [44] proposed the **Piecewise Mechanism (PM)** that combines the merits of the Laplace mechanism and Randomized response mechanism. It confines the perturbed value to a relatively small domain and allows it to be adjacent to the real value with rational probability.

We first introduce the PM for single numerical attribute data. The PM takes as input a value $t_i \in [-1, 1]$ and returns a perturbed value $v_i^* \in \{-C, C\}$. v_i^* is calculated by the following distribution:

$$\Pr[v_i^* = a] = \begin{cases} \dfrac{e^{\varepsilon} - e^{\varepsilon/2}}{2 \cdot \left(e^{\varepsilon/2} + 1\right)}, & \text{if } a \in [l(v_i), r(v_i)], \\ \dfrac{e^{\varepsilon} - e^{\varepsilon/2}}{2e^{\varepsilon} \cdot \left(e^{\varepsilon/2} + 1\right)}, & \text{if } a \in [-C, l(v_i)) \cup (r(v_i), C)], \end{cases} \tag{37}$$

where $C = (e^{\varepsilon/2} + 1)/(e^{\varepsilon/2} - 1)$, and $l(v_i) = 2^{-1} \cdot (C + 1) \cdot v_i - 2^{-1} \cdot (C - 1), r(v_i) = l(v_i) + C - 1$.

Given each participant's exact vector $t_i \in [-1, 1]^d$, PM-based Mean estimation for multiple numerical attributes returns a perturbed vector $t_i^* \in [-d \cdot C, d \cdot C]^d$. Similar with the Harmony-mean method, PM first initializes the perturbed vector $t_i^* = [0]^d$ and sets to be $k = \max\{1, \min\{d, \lfloor 2\varepsilon/3 \rfloor\}\}$, then selects a set $\boldsymbol{K}$ with size k randomly without replacement from $[d]$, next loops through the set $\boldsymbol{K}$ and feeds and into PM, and obtains the noise value $x_{i,j}$ and the perturbed value of the corresponding attribute $t_i^*[j]$ is $d \cdot x_{i,j}/k$. The PM achieves $O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$ asymptotic error for the setting of multiple attributes.

Since PM can still be slightly worse than MeanEst in the worst-case variance when the privacy budget is less than 1.29, Wang and Xiao et al. [44] further proposed **Hybrid Mechanism (HM)** to maintain the advantages of PM by combining PM and MeanEst. We first introduce the Hybrid Mechanism that is proposed for single numerical attribute data. Specifically, HM takes as input value $t_i \in [-1, 1]$, and it then flips a coin with head probability α; if the coin is a head (resp. tail), then it calls PM (resp. MeanEst) to randomized t_i. The noise variance generated by HM is

$$\sigma_H^2(t_i, \varepsilon) = \alpha \cdot \sigma_P^2(t_i, \varepsilon) + (1 - \alpha) \cdot \sigma_M^2(t_i, \varepsilon), \tag{38}$$

where $\sigma_P^2(t_i, \varepsilon)$ and $\sigma_P^2(t_i, \varepsilon)$ denote the noise variance generated by PM and MeanEst, respectively. The setting and analysis of optimal parameter α have presented in [44]. HM-based Mean estimation for multiple numerical attributes' implement is similar to the PM-based method.

The HM can achieve optimal result utility (worse-case noise variance) for mean estimation on the multiple numerical attributes compared to existing methods according to [10, 44]. It can obtain an asymptotic optimal error $O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$. Table 5 shows the fundamental methods of mean estimation for multiple attributes data under LDP. Note that A is determined by ε and d, that is equal to $1 + \mathrm{Lap}(2d/\varepsilon)$.

Recently, there are some existing studies for extending the above methods for LDP mean estimation. Wang et al. [49] extended the pure ε LDP to the approximate (ε, δ)-LDP and designed several (ε, δ)-LDP algorithms for collecting multidimensional numerical attribute data. These algorithms provide higher accuracy than the optimal Laplace mechanism while guaranteeing the privacy for each user. Since these algorithms are the variants of the above-mentioned method, we will not go into them in detail. Akter et al. [50] borrowed the definition of personalized local differential privacy (PLDP) [51] and adopted the Harmony-mean method for mean estimation over numerical data. Li et al. [52] studied the problem of locally differentially private distribution estimation on numerical attribute data. They firstly proposed to apply the Alternating Direction Method of Multipliers optimization to post-processing Hierarchical Histograms (HH) [38] for estimation improvement, which is called HH-ADMM method and is regarded as Categorical Frequency Oracles to reconstruct the numerical distribution by discretizing numerical domain into the categorical domain. Since the method does not fully exploit the numerical nature, they proposed a square wave (SW) mechanism and combined its reporting with Expectation Maximization and Smoothing (EMS) to improve the result of estimation.

Existing studies focus on the theoretical analysis of mean estimation under LDP. Smith et al. [53] studied the problem of high-dimensional mean estimation under noninteractive (ε, δ)-LDP and found that it can be achieved with logarithmic dependence on dimensionality by adopting random projection and approximate techniques under the assumption on data points within ℓ_2-norm bound. Gaboardi et al. [54] investigated the bound problem of mean estimation under the (ε, δ)-LDP and provided tight upper and lower bounds for the problem assuming the each participant's data are drawn from an unknown Gaussian distribution (mean or variance is unknown). Likewise, Joseph et al. [55] further studied the problem and presented a smaller lower bound of mean estimation in the LDP than [54].

5.2. Locally Differentially Private Distribution Estimation

5.2.1. Distribution Estimation Methods in Server-Side. The previous methods in Section 4 focus on the adoption of different user-side private mechanisms and basic frequency statistics approach (straightforward statistic-based approach, counting) in the server-side estimation methods. We will further discuss the several server-side methods for distribution estimation methods. Existing LDP distribution estimation methods in server-side mainly include the Empirical estimation method (Matrix inversion method) and EM reconstruction method.

The locally private distribution estimation problem: Given a distribution vector $\mathbf{p} = (p_1, \ldots, p_k)$ on the probability simplex $\mathbf{S}^k$, samples $X_1, \ldots, X_n$ are drawn i.i.d. according to $\mathbf{p}$. ε-LDP mechanism is independently applied to each sample X_i to generate the private observations. Our goal is to estimate the distribution vector $\mathbf{p}$ from $\mathbf{Z^n}$.

(1) Empirical estimation method (matrix inversion method)

The empirical estimation method [35, 56] is also regarded as matrix inversion method. The empirical estimate $\widetilde{\mathbf{p}}$ is computed by using an empirical distribution $\widetilde{\mathbf{q}}$ of the perturbed data $\mathbf{Z}$. Note that $\widetilde{\mathbf{p}}, \widetilde{\mathbf{q}}$, and $\mathbf{M}$ are denoted as an input domain $|\mathbb{X}|$ dimensional vector, its corresponding output domain

Table 5: Fundamental methods of mean estimation for multiple numerical attributes data under LDP.

Methods	Input boundary	Output boundary	Communication cost	Asymptotic bound error	Privacy budget/number of attributes
LM [20]	$[-1,1]$	$[-A,A]$	$O(d)$	$O(2d/\varepsilon\sqrt{n})$	ε/d
SCDF LM [47]	$[-1,1]$	$(-\infty,\infty)$	$O(d)$	$O(2de^{\varepsilon/2}/(e^{\varepsilon}-1)\sqrt{n})$	ε/d
Staircase mechanism [48]	$[-1,1]$	$(-\infty,\infty)$	$O(d)$	$O(2de^{\varepsilon/2}/(e^{\varepsilon}-1)\sqrt{n})$	ε/d
MeanEst [30]	$[-1,1]$	$\{-B,B\}$	$\Theta(d)$	$O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$	ε/d
Harmony-mean [10]	$[-1,1]$	$\{-B,0,B\}$	$\Theta(1)$	$O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$	ε
Piecewise mechanism [44]	$[-1,1]$	$[-dC,dC]$	$O(k)$	$O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$	ε/k
Hybrid mechanism [44]	$[-1,1]$	$[-dC,dC]$	$O(k)$	$O(\sqrt{d\log(d)}/\varepsilon\sqrt{n})$	ε/k

$|\mathbb{Z}|$ -dimensional vector, and $|\mathbb{X}|\times|\mathbb{Z}|$ matrix, respectively. Their relationship is formally shown by the following equation:

$$\widetilde{\mathbf{p}}\cdot\mathbf{M}=\widetilde{\mathbf{q}}. \tag{39}$$

Computing the $\widetilde{\mathbf{p}}$ is done by solving the above equation. The empirical distribution $\widetilde{\mathbf{q}}$ can converge to the exact distribution $\mathbf{q}$ only when the number of participants is relatively larger. Hence, the empirical distribution$\widetilde{\mathbf{p}}$ can converge to the distribution $\mathbf{p}$.

A main disadvantage of the method is that some elements in $\widetilde{\mathbf{p}}$ may be negative due to the small number of participants. Two methods are proposed to fix it by Kairouz et al. [35], one method is a *normalized decoder method* that truncates the negative elements of $\widetilde{\mathbf{p}}$ to 0 and renormalizes $\widetilde{\mathbf{p}}$ so that its sum is 1. Another method is a *projected decoder method* that projects onto the probability simplex $\boldsymbol{C}$ so that the Euclidean distance between any two points is minimized.

(2) EM reconstruction method

Since the Expectation-Maximization (EM) algorithm is a general method to approximate maximum likelihood estimates (MLE) of unknown parameters in the presence of missing or incomplete data (e.g., the case of a small number of participants), the EM-based reconstruction method (also regarded as the iterative Bayesian method) is proposed to deal with it. In particular, the EM reconstruction method works on the perturbed data $\mathbf{Z}$ by fixing iteratively the parameter $\widetilde{\mathbf{p}}$ so that the expectation of the log-likelihood function $L_{\mathbf{Y}}(\widetilde{\mathbf{p}})=\log(\Pr(Z\,|\,\widetilde{\mathbf{p}}))$ is maximized. Therefore, the method's feature is that the converged estimate result is equivalent to the maximum likelihood estimate in the probability simplex without considering the number of participants.

For example, Fanti et al. [57] introduced a novel algorithm (called RAPPOR-unknown) for distribution estimation of string-valued data with unknowing the possible domain (set of possible values) in advance. Specifically, each participant firstly adopts the Basic RAPPOR to perturb the multiple substrings (n-grams) from his string-value data and sends the multiple perturbed substrings to the collector. Then the collector learns the distribution estimation of string-value data by constructing the joint distributions of all possible substrings using an EM-based algorithm. The algorithm can be generalized to other LDP algorithms that learn a distribution of discrete string-valued data.

In the above methods, there exists an assumption that the server knows the obfuscation mechanism Q that is adopted by each participant and is the same for them (symmetric scheme). In addition, Ye et al. [58] investigated the problem of locally differentially private distribution (frequency) estimation in multiple regimes (levels) scenarios where each participant who has a personalized privacy budget ε pertains to a group with certain privacy regime. They formulated the problem and proposed several Maximum likelihood estimation (MLE) methods based on group operation to handle the different situations (asymmetric scheme) when participants' privacy regimes are in some practical cases.

5.2.2. Distribution Estimation with Small Sample Problem and Linear Queries Estimation. We further introduce some special problems of distribution estimation under LDP, which include small sample problem and linear queries estimation for distribution estimation.

(1) Small Sample Problem. Sei et al. [43] proposed two novel locally private distribution estimation schemes for anonymized data collecting, namely, Single to Randomized Multiple Dummies (S2M) and S2M with Bayes (S2Mb). The basic block of S2M is the k-Subset mechanism used in the user-side, and the EM (Expectation-Maximization) algorithm is adopted for reconstructing the distribution. Both schemes materialize anonymization and reconstruct data distribution more accurately by generating a set of disguised values from one real value. The schemes are suitable for distribution estimation with small samples while guaranteeing LDP. Murakami et al. [59] studied on the locally

differentially private distribution estimation with small samples (the number of participants n is small). They first analyzed two statistical inference methods, matrix inversion method and EM (Expectation-Maximization) reconstruction method, and showed that the latter outperforms significantly the former in terms of utility (estimation error) and then proposed a method to correct estimation error of EM reconstruction based on Rilstone's theory. Gursoy et al. [60] introduced a notion of Condensed Local Differential Privacy (CLDP) to solve the small user population (small samples) problem and develop a set of CLDP protocols for privacy-preserving differential types data, ranging from ordinal items to nonordinal items and to sequences of ordinal and nonordinal items, while providing desirable statistical utility.

(2) Linear Queries Distribution Estimation. Bassily [61] focused on the study of locally differentially private linear queries estimation which conducts a batch of d linear queries on certain unknown distribution and includes various estimation problems such as distribution estimation and d-dimension mean estimation. The work provided several algorithms for this problem under both interactive setting and noninteractive setting. In the noninteractive setting, the batch of d queries is expressed as the rows of matrix $\mathbf{M} \in \mathbb{R}^{d\times|D|}$ that is generated by the server before the protocol begins. In the interactive setting, the batch of d queries is conducted over d rounds, namely, one query in each round. Note that the above-mentioned distribution estimation and mean estimation over a finite domain can be regarded as the special offline (noninteractive) queries. For the high-dimensional linear queries (number of queries, e.g. $d \geq n$), the work presented a new noninteractive algorithm with the L_2 error that is independent of d in the high-dimensional setting and is sublogarithmically sustained by the domain size. Besides, the work also provided a novel interactive algorithm with optimal L_∞ (maximum) estimation error, where the upper bound in the interactive setting is equivalent to the lower bound in the nonadaptive setting in terms of L_∞ error.

5.2.3. Joint Distribution Estimation over Multivariate Data. There are some studies on the joint distribution estimation over multivariate data. RAPPOR-unknown [57] is a methodology for estimating the joint distribution of multiple variables (Multivariate). To estimate the joint distribution and perform a formal statistical test for independence, the method combines EM algorithm with the variance-covariance matrix. The former is applied to estimate the joint distribution. The latter is applied to enable testing for association. Cormode et al. [62] studied the marginal release (joint distribution) under LDP, provided a set of algorithms for achieving marginal statistics with the stronger model of LDP and the tight theoretical bounds on the utility of marginal statistics, and performed empirical evaluation about these bounds. Zhang et al. [63] proposed a novel Consistent Adaptive Local Marginal (CALM) method for computing any k-way marginal (joint distribution) under the local setting of differential privacy.

Peng et al. [64] proposed a locally differentially private joint distribution method for location and assigned sensing attributes (environmental attributes, e.g., air quality) in crowdsensing scenarios. They proposed an optimized LDP algorithm that combines the advantage of k-Subset mechanism and RAPPOR mechanism to achieve the local data protection. Ren et al. [65] investigated the locally private joint distribution estimation from the large-scale and high-dimensional crowdsourced data and proposed an EM-based joint (multivariate) distribution estimation over high-dimensional data. However, the performance of EM-based method is poor since the method needs to scan all participants' data. They further adopted Lasso Regression for the distribution estimation to improve the performance, namely, reduction of computation complexity. Note that each participant uses the one-time RAPPOR mechanism (RAPPOR without IRR) to locally perturb their original data.

Soon after, building on the EM and Lasso Regression-based joint distribution estimation for multidimensional data, Ren's team [66] further developed a dimensionality reduction method based on an undirected dependency graph that captures (identifies) the correlated attributes in the high-dimensional data and proposed a **Lo**cally differentially private data **Pub**lication (LoPub) scheme for the high-dimensional crowdsourced data.

5.3. Private Estimation over Other Complex Data with LDP

5.3.1. Itemset Related Data. Several studies have focused on the locally differentially private **Set-valued data**, which is a set of items. Qin et al. [67] focused on the study of heavy hitter estimation over the set-valued data with LDP and proposed a two-phase framework called LDMiner for dealing with the issue. Wang et al. [68] investigated the locally differentially private frequent itemset discovery over the set-valued data. Wang et al. [69] proposed an efficient and effective local differential private set-valued aggregation mechanism called PrivSet. **Key-value data** is a well-popular NoSQL data model and a generalized form of set-valued and numerical data. Ye et al. [70] focused on locally differentially private key-value data and proposed several LDP-preserving solutions for frequency estimation and mean estimation on key-value data, namely, PrivKV and its two iterative improvements in terms of estimation accuracy. In addition, Yang et al. [71] studied the problem of locally differentially private collection of **Preference Rankings**, which puts unequal items into a total order depending on personal opinions on their relative quality and proposed a novel approach called Sampling Randomizer For Multiple Attributes with Riffle Independent Model (SAFARI). It adopted the riffle independent model to collect a group of distributions over small domains to estimate the overall distribution of users' rankings and used the collected distributions to construct a synthetic ranking dataset. Yan and Li et al. [72] investigated the problem of private ranking aggregation under LDP and proposed a LDPKwikSort protocol for achieving it with acceptable utility of aggregation estimation.

5.3.2. Ordinal Data. **Ordinal data** are the categorical data with linear ordering among categories, including the discrete numerical data (such as discrete sensor data) and other categorical data (such as preference options). Some studies focused on the investigation of locally differentially private distribution estimation on the original data. In [73] and its extended work [74], Wang et al. proposed an efficient and effective locally private mechanism for the problem, namely, Subset Exponential mechanism (SEM), which is an extension of k-subset mechanism that randomly responds with fixed-size subset with designed probability and further proposed Circle Subset Exponential mechanism (CSEM) based on the circling distance technique for special uniform ordinal data so that the complexity of computation and space are reduced and the theoretical error bounds are tight. However, similar with k-subset mechanism, the SEM and CSEM will bring more communication overhead.

5.3.3. Text Structure Data. Private collecting text data is an important application of LDP. Most studies (based on the mechanisms in Section 4) focused on the learning frequency of word with known text (word) domain, but new word problem which is regarded as the frequency estimation with the unknown domain is common in the application. The O-RR and O-RAPPOR were designed for the open domain [35]. However, they are unsuitable for the problem of learning new words, since the set of candidate words is extremely large. Fanti et al. [57] proposed a novel method for estimating the frequencies of unknown strings. Its main idea is to utilize concurrences among n-grams to reconstruct a set of candidate words by EM technique. Thakurta et al. [75] proposed an LDP collecting new words method for Apple Inc, which needs to collect two LDP reports for a single word from each participant, namely report of a single word and report of a single n-gram. The LDP report of n-gram consists of the hash value of the word and the selected n-gram. Different from RAPPOR-Unknown and Sketch method that requires multiple LDP reports for a single word, Kim et al. [76] proposed a novel LDP method for privately collecting new words by generating only one report for a single word to improve the efficient use of privacy budget and reduce the computational cost with the help of the idea from message authentication, where each user sends a noisy report of one n-gram selected randomly from a single string consisting of a new word and its hash value. The server then decodes the collected reports of n-grams, discovers new words using the links between partially overlapping n-grams, and checks integrity with hash values. In addition, different from the above n-gram based methods, Wang and Xiao et al. [77] proposed a trie-based method for the problem of learning new words, called PrivTrie. However, the method leads to more computation and communication overhead due to the iterative construction of the trie and multiple rounds of interaction between a participant and the server.

5.3.4. Graph Structure Data. There are many types of complex relations in the sensitive individual data. The graph is used to model these relations. For example, relations between users are represented into a simple graph, or relations between users and other entities are represented into a bipartite graph. To protect the privacy of these data, Qin et al. [78] proposed LDPGen, a novel and effective multiphase method to gradually extract information from users and build a relational graph of the fundamental social network, while preserving the edge local differentially privacy. Note that LDPGen is directly building on the existing LDP and synthetic graph generation techniques. Zhang et al. [79] also focused on the same problem and proposed an optimized randomized response for private synthetic graph generation with LDP. Gao et al. [80] transformed the subgraphs into neighbor profiles of the HRG (hierarchical random graph) and injected noise into the neighbor profiles to achieve the LDP. Considering the privacy concern of obtaining the participant's privacy information from their neighbors in social networks and ensuring not only their privacy but also the privacy of their neighbors, Sun and Xiao et al. [81] proposed a definition of decentralized differential privacy (DDP, which is a special LDP with the same privacy budget for each participant) for social graph analysis and designed a novel framework for estimating accurately the subgraph counts in the global graph with DDP by adopting the local graph structure. Recently, Wei et al. [82] proposed a novel AsgLDP method to generate privacy-preserving attributed graph data under LDP, its advantage is that it can protect various graph properties (e.g., degree distribution, community structure, and attribute distribution) with LDP and achieve a superior tradeoff between utility and privacy.

5.4. Private Estimation over Streaming (Evolving) Data with LDP. The problem of collecting user statistics across time periods is more practical. Recently, there are several studies on collecting evolving data under LDP. RAPPOR [7] is originally proposed to collect the chrome's profiles data over time under the LDP. The longitudinal privacy is guaranteed by using the permanent random response (PRR) that is seen as a memoization technique; however, RAPPOR is only used for private evolving data that will not change over time due to the adoption of PRR. Ding et al. [9] developed several novel LDP mechanisms based on memoization for continual collection of counter data, and its privacy guarantee does not degrade over time. These mechanisms have been deployed in millions of devices running Windows 10 OS by Microsoft to collect application usage statistics while holding users' privacy.

The above two methods have adopted the memoization technique to deal with the constant or very small-but-frequent change value problem over time, but memoization causes the storage overhead for each participant's device. Zhao and Chen et al. [83] proposed the SAnonLDP algorithm by combining k-anonymity and LDP, which subsumes four basic modules: random grouping; anonymous and Walsh–Fourier transforms; random response; and SVD (singular value decomposition). The main idea of the algorithm is adopting the Walsh Fourier transform that integrates k-anonymity into LDP to reduce the memory and communication overhead while guaranteeing the acceptable frequency estimation by other modules.

Joseph et al. [84] proposed a novel LDP technique for distribution estimation to preserve up-to-date statistics over time, with privacy guarantees that degrade only with the number of distribution changes rather than the number of periods. To deal with it, they proposed a thresh algorithm and consensus voting protocol. The former's main idea is to update the global estimate only when it might become sufficiently inaccurate and thus take advantage of the possibly small number of changes in the underlying statistic. The latter is used to identify "update needed" epochs. The participants will check their data and privately release a vote for whether the global estimate needs to be updated. The consensus voting process is equivalent to the locally differentially private heavy hitter identification problem. However, the method will bring about more communication overhead due to multiple round interactivities.

In addition, Bittau et al. [85] proposed the Encode, Shuffle, Analyze (ESA) architecture for monitoring with high utility while preserving user privacy. It is the first systematic architecture for privacy-preserving software monitoring depending on cryptography, anonymity, and differential privacy. Erlingsson et al. [86] further studied and derived that the combination of LDP and anonymity (via shuffling) can provide stronger differential privacy bounds, and presented a real-time monitoring protocol for guaranteeing longitudinal privacy of users' report over timestamps, irrespective of whether their report is about independent or correlated values. However, these methods are based on the shuffle model that adds a shuffler into LDP; this will result in more complexity of system model.

6. Current Research Circumstances

6.1. Private Statistical Learning (Inferencing) under LDP. There are many studies focused on applying LDP to private statistical learning and inferencing. We mainly give an overview of private empirical risk minimization, private hypothesis testing, and private federated learning and deep leaning.

6.1.1. Private Empirical Risk Minimization under LDP. Kasiviswanathan et al. [5] first investigated the learning problem under LDP and presented a general balance between learning of LDP and that of statistical query model. Duchi et al. [87] studied the statistical risk minimization problem under LDP by computing saddle points of mutual information and exhibited a precise tradeoff between the privacy and the utility measured by convergence rate of any statistical estimator or learning procedure. However, the work required the participants should be optimally private and could communicate merely by transferring a perturbed gradient of the selected loss function. Duchi et al. [30, 31] proposed a formal framework, which is dependent on the classical minimax risk for characterizing the tradeoff between utility and LDP. Its main goals are to characterize how the optimal estimation rate varies with the privacy level and other problem parameters for various estimation problems. They developed private versions of classical information-theoretical bounds based on the Le Cam, Fano, and Assouad inequalities which achieve precise minimax rates under local privacy constraints and develop provably (minimax) optimal estimators. Moreover, they provided the lower bounds and optimal mechanisms for general convex optimizations, but these optimal procedures need more rounds of interactions. Smith et al. [53] first studied some convex optimization problems under LDP including the interactive LDP problem of round complexity and the noninteractive LDP problem with general convex loss functions. They discovered that the structure of an optimization problem affects not only the accuracy but also the amount of interaction that is necessary to get that accuracy.

Zheng et al. [88] proposed efficient algorithms for general learning problems under noninteractive LDP and demonstrated the dependence of excess risk in the case of high dimension (e.g., sparse linear regression and kernel ridge regression). Wang et al. [89] studied the ERM problem in the noninteractive LDP model and showed that if the loss function is (∞, T)-smooth, the error bound does not depend on the sample complexity in the case of constant or low dimensionality $n >> p$; if the function is a linear convex function, its error bound relies on the Gaussian width of the underlying constrained set, not affected by p in the case of high dimension (sparse case, e.g. Sparse Linear Regression) $n << p$. Wang et al. [90] proposed a general ERM approach under noninteractive LDP by using a polynomial of inner product approximation rather than directly using the polynomial approach in their previous work [89]. Wang and Xu [91] revealed that the polynomial dependency on the dimensionality p is inevitable for the estimation error in noninteractive and sequential interactive settings of LDP and proposed a sequential interactive LDP algorithm for the low-dimensional sparse case that can be used to settle LDP-ERM with sparsity constraints and sparse nonlinear regression and a general algorithm for a restricted high-dimensional sparse case. The above-mentioned studies have focused on the theoretical investigation about ERM. Besides, ERM also attracts some studies from a practical perspective (e.g., Nguyên et al. [10] and Wang et al. [44]).

In addition, Feldman et al. [92] focused on the study on private learning problem and demonstrated that contractive iterations can strongly amplify the privacy preservation under the various variants of differential privacy without depending on publishing intermediate results that are commonly adopted in the previous differentially private learning algorithms. Moreover, they proposed an Online Convex Optimization framework that is used to design and analyze the algorithms for training machine learning models. Hoeven et al. [93] extended LDP to the unconstrained Online Convex optimization learning and provided personalized privacy preservation for data providers. Table 6 shows the existing methods of private ERM under LDP.

6.1.2. Private Hypothesis Testing under LDP. Hypothesis testing is a commonly used statistical inference tool. There are also some studies on private hypothesis testing under

Table 6: Existing methods of private ERM under LDP. Bounds errors and complexity for optimization of convex functions in the local model as a function of the number T of rounds interaction (in the case of interactive setting), the number n of participants, and the dimension p of the parameter vector. α is the desired population excess risk (expected empirical error), optimization error is equivalent to expected population risk.

Method	Dimension (method)	Interactivity	Problem	Assumption on loss function	Assumption on variables x_i, y_i and constrain set C	Bound error/complexity
Duchi et al. [31]	Low	Seq.	General convex optimization (minimax risk)	Generalized convex loss	$\|C\|_2 \leq 1$	Optimization error lower bound $O(\sqrt{p/n\varepsilon^2})$
Smith et al. [53]	Low	Non.	General convex optimization	Generalized convex loss	$\|C\|_2 \leq 1$	Optimization error $O((\sqrt{p}/n\varepsilon^2)^{1/(p+1)})$ sample complexity: $\Theta(p(\varepsilon\alpha)^{-2})$ (linear regression)
	Low	Seq.		Lipschitz convex loss		Round complexity: $\Theta(\log(1/\alpha))$ Optimization error: $O(\sqrt{T}^{(-1)} + \sqrt{p/n\varepsilon^2})$
Zheng et al. [88]	High (dimension reduction)	Non	Linear regression	Sparse linear regression	$\|x_i\|_2 \leq 1$ $\|y_i\|_1 \leq 1$ $\|C\|_1 \leq 1$	Optimization error: $O(\sqrt[4]{\log p/n\varepsilon^2})$
	High (polynomial approximation)	Non		Smooth generalized linear		Sample complexity $p \cdot (1/\alpha)^{O(\log\log(1/\alpha)+\log(1/\varepsilon))}$
Wang et al. [89]	Low	Non	Convex optimization with sample complexity	$(8, T)$-Smooth	$\|x_i\|_2 \leq 1$ $\|y_i\|_1 \leq 1$ $\|C\|_1 \leq 1$	Sample complexity $O((cp^{1/4})^p a^{-(2+p/2)} \varepsilon^{-2})$
Wang et al. [89]	Low (polynomial approximation)	Non	Convex optimization with sample complexity	(∞, T)-Smooth		Sample complexity $O(4^{p^2+p} D_p^2 \varepsilon^{-2} a^{-4})$
Wang et al. [89]	High (dimension reduction)	Non	Convex optimization	Smooth generalized linear		Optimization error $O(\sqrt[4]{\log p \log n}/(\sqrt{n}\,\varepsilon))$
Wang et al. [90]	High (polynomial approximation)	Non	Convex optimization (ERM)	Lipschitz convex generalized linear loss	$\|x_i\|_2 \leq 1$ $\|y_i\|_2 \leq 1$ $\|C\|_1 \leq 1$	Sample complexity
Wang and Xu [91]	High (polynomial approximation)	Seq.	Sparse linear regression	Squared L_2 loss	$\|x_i\|_2 = \sqrt{p}$ $\|y_i\|_1 \leq 1$ $\|C\|_1 \leq 1$	Estimation error lower: $\Theta(\sqrt{p/n\varepsilon^2})$ upper: $O(\sqrt{p \log p \log n/(n\varepsilon^2)})$

LDP. Gaboardi et al. [94] explored the design of private hypothesis tests in the LDP model and analyzed locally private chi-square testing and independence testing. Sheffet [95] focused on the study of differentially private hypothesis testing in the local model under symmetric (same randomized function) and asymmetric (personalized randomized function) randomized response mechanism and investigated the general framework of mapping each user's type into a signal and then provided sample complexity bounds for identity (uniformity) and independence testing under randomized response.

Acharya et al. [96] used several already deployed general LDP mechanisms (RAPPOR and Hadamard Response) and proposed a bespoke mechanism (Randomized Aggregated Private Testing Optimal Response, RAPTOR) for testing, and proposed identity and independence testing based on the above mechanisms and analyzed their sample complexities. The analysis results show the proposed testing algorithm based on Raptor is sample-optimal. Specifically, the testing based on Raptor requires significantly fewer samples than any testing based on RAPPOR [7] or Hadamard Response [39].

Canonne et al. [97] investigated the problem of optimal sample complexity about private hypotheses testing. In particular, given two distributions P and Q and the privacy level ε, they characterized a sample complexity up to constant factors in terms of the above-given conditions, which is achieved by certain randomized hypotheses testing. The result is parallel to the classical Neyman–Pearson lemma in the setting of private hypothesis testing and extends into the private change-point detection application. Joseph et al. [33] studied the problem in different interactive settings and showed that for any simple hypothesis testing and compound hypothesis testing with convex and compact distributions, noninteractive LDP protocol can achieve optimal sample complexity. They demonstrated a simple hypothesis testing which is optimal among all full interactive tests by using the information-theoretical lower bound techniques.

In addition, Gaboardi et al. [54] adopted the private Z-test to locally private mean estimation in the known variance of Gaussian distribution $\mathbb{N} \sim (u, \sigma^2)$. Particularly, an aggregator (analyst) can estimate confidence interval based on the random sample data from the Gaussian distribution and then reject (or fail-to-reject) the parameter u's certain hypotheses. For example, the hypothesis is that the means of two independent sample sets are equal. The aggregator can estimate the mean of the Gaussian distribution by the private Z-test.

6.1.3. Private Federated Learning and Deep Learning in the Local Model

(1) Federated Learning. Geyer et al. [98] proposed an algorithm for LDP federated learning to protect the personal data by hiding the participant's contributions during training while guaranteeing the high performance of model. McMahan et al. [99] introduced an algorithm for locally differentially private training of complex sequence neural network model for language predict and proposed the first locally differentially private training LSTM (long short-term memory) based language model trained without significant accuracy loss of model. Li et al. [100] proposed several locally differentially private meta-learning algorithm that can be adopted for federated learning. In [101], Bhowmick and Duchi et al. described the analysis and implementation with novel optimal LDP mechanisms for federated learning system in the setting of curious adversaries; they introduced relaxed local privacy by constraining the adversaries' power and proposed a two-pronged approach for locally differentially private Federated Learning with the smaller privacy parameter ε and obtained high-performance models that are near nonprivate approach. Note that LDP with smaller ε makes the model-fitting process extraordinarily challenging.

(2) Deep Learning. Arachchige et al. [102] proposed a novel LDP mechanism named LATNET to train a deep neural network (DNN) with high privacy and accuracy. LATNET redesigns the training process by splitting the CNN architecture into the convolutional layer, randomization layer, and fully connected layer, which randomization layer is provided privacy preservation by LDP. Xu et al. [103] proposed EdgeSanitizer, an edge computing oriented deep inference framework with LDP for mobile data analytics. Specifically, the framework leverages a deep learning-based data minimization model to constrain the data size and obfuscates the learned features from original data by adaptively injecting noise to achieve the LDP, hence constructing a novel privacy preservation layer against sensitive information inference on the edge server.

6.2. Private Statistical Data Analysis under LDP

6.2.1. Private Correlation and Clustering Analysis under LDP

(1) Private Covariance Matrix Estimation. Wang et al. [104] focused on the study of sparse covariance matrix estimation problem under the differential privacy, proposed a novel DP-Thresholding method to achieve the ℓ_2-norm error bound, and further extended the bound to a general-norm based one $(1 \leq w \leq \infty)$. The method is significantly better than the one adding noise directly and easily extended to the LDP model. Wang et al. [105] moved forward to study the problem of sparse covariance matrix estimation under the LDP model, gave a lower bound on the noninteractive private minimax risk by using the measurement of squared spectral norm, and proposed a General Private Assouad Lemma framework for bounding the private minimax risk of matrix-related estimation problems.

(2) Private Principal Component Analysis. Balcan et al. [106] first studied the private PCA problem under the LDP model and provided an improved noisy power method (Laplace mechanism) for the general setting. It is worth noting that the output is only $O(k)$-dimensional subspace, rather than the exact k-dimensional subspace. Ge et al. [107] investigated the PCA for high-dimensional data under the LDP model and proposed a locally differentially private sparse PCA algorithm with the optimal minimax statistical error. Wang et al. [108] investigated the PCA problem under the noninteractive LDP model from the perspective of low-dimensional and high-dimensional (row sparse) cases, respectively. For the low-dimensional case, they showed the optimal rate for the private minimax risk of the k-dimensional PCA by measuring squared subspace distance. For the high-dimensional case, they provided an efficient algorithm to achieve an approximate optimal upper bound.

(3) Private Clustering Analysis. Nissim et al. [109] first proposed local differential private algorithm LDP-GOODCenter for the k-means clustering by combining local sensitive hashing technique with the heavy hitter algorithm [110] to discern a set of points falling into an approximate smallest closed ball. They then computed the approximate average of these discerned points under LDP. However, it only researched the 1-cluster in the local model. Kaplan et al. [111] designed a set of novel differentially private Euclidean k-means algorithms under the local model. The algorithms outperform the previous algorithms in the context of multiplicative error and significantly reduce the number of interactions.

6.2.2. Data Mining Analysis under LDP

(1) Private Heavy Hitter Identification (Frequent Item Mining). The goal is to identify the items that are frequent. When the size of the possible domain is small, this can be solved with an FO protocol. If the possible domain is very

large, this way is computationally infeasible. There are several related studies in the LDP. Hsu et al. [112] and Mishra et al. [113] studied the problem of private heavy hitter with LDP and provided efficient protocols for the problem. However, their error bound is higher than the SHist protocol proposed by Bassily and Smith et al. [37]. Bassily et al. [110] further proposed practical local differentially private heavy hitters algorithm called TreeHist with higher accuracy and lower time complexity than SHist, which is their previous work. Acharya et al. [114] investigated the tradeoffs between accuracy (utility) and communication complexity for locally differentially private heavy hitter estimation and theoretically demonstrated that the Hadamard Response is utility-optimal for heavy hitter estimation. In addition, the lower bound of communication complexity of the optimal heavy hitter estimation algorithm without public randomness was given. Bun et al. [115] proposed a locally differentially private heavy hitter algorithm called PrivateExpanderSketch with cutting-edge theoretical performance as a function of all standardly given parameters. Wang et al. [116] proposed a locally private heavy hitter method with large domains named Prefix Extending Method (PEM). The method's fundamental idea is to iteratively identify gradually longer frequent prefixes. Jia et al. [45] also studied the problem of locally private heavy hitters by incorporating the prior knowledge of noise item frequencies and exact item frequencies to reduce estimation errors.

(2) Private Frequent Itemset Mining. Compared to the heavy hitter identification, the problem is a more practical setting where each participant's value is a set of items drawn from the item domain. Many studies applied LDP to frequent Itemset mining. For example, Apple applied LDP to privately collect the frequency estimation of the emojis typed by users to improve the functions, each user has a set of emojis that they typed [8]. The problem is quite challenging. Encoding the original domain into the private domain (power set of the original domain) and using existing FO protocol do not work.

To settle the issue, Qin et al. [67] proposed a two-phase framework based on LDP called LDPMiner to discover top-k heavy hitters over set-valued data. The framework requires one or two rounds of interaction between each participant and the data collector. Its main idea is to select $O(k)$ candidate heavy hitters in the first phase and then compute the approximate frequency of these candidates in the second phase. LDPMiner has lower communication overhead between participants and data collector than the existing frequent itemset mining algorithms. Wang et al. [68] designed a locally differentially private frequent itemset mining protocol that obtains better in the context of accuracy than LDPMiner under the same privacy budget. The merit benefits from the crucial observation, namely, privacy amplification with sampling, and an optimal FO algorithm, that is Optimal Local Hash [36]. Zhang et al. [117] proposed M-RR mechanism for locally differentially private FIM based on k-RAPPOR and k-RR. The main idea of M-RR is to divide the privacy budget into ε_s and ε_b in the light of participants' privacy concern levels, then apply k-RAPPOR with ε_s and k-RR with ε_b to perturb the low and high attributes, respectively. Wang and Xiao et al. [77] proposed PrivTrie for estimating frequencies of candidate terms (similar with itemset) by iteratively building structure trie under LDP, different from using n-grams in the RAPPOR-unknown [57]. PrivTrie adopted a novel LDP-compliant trie-building algorithm to assign most of privacy budget on the frequencies estimation of exact terms and achieve high utility.

6.3. Privacy Amplification for LDP. The techniques of privacy amplification for LDP include anonymity, shuffling, and sampling-iteration in the LDP. LDP model where the server or aggregator collecting private data from participants can know any specific participant by retracing their input data can only guarantee the protection of data privacy, but not the protection of identity privacy. To solve the shortcoming, Zhao et al. [83] applied k-anonymous technique in locally differentially private frequency estimation algorithm to achieve the privacy amplification under in LDP model. However, the method is a simple combination of k-anonymous and LDP; it cannot defend the background knowledge attack.

Recent, Bittau et al. [85] proposed the Encode, Shuffle, Analyze (ESA) architecture for monitoring with high utility while preserving data privacy and identity privacy to achieve privacy amplification. The core idea of ESA is the shuttle model, which is LDP with the anonymity (shuttling) technique used for the post-processing process of LDP. The comparison of local and shuttle models is shown in Figure 3. The shuttle model adds a Shuffler module between participants (Abbreviated as particip.) and the data collector (server), compared to the local model. The Shuffler module is used to shuttle the private data $Z_1, Z_2, \ldots, Z_n$ into $S(Z_1, Z_2, \ldots, Z_n)$ so that the collector cannot identify any specific participant by tracking back their input data.

Erlingsson et al. [86] showed that the combination of LDP and anonymity (via shuffling) can provide stronger differential privacy bounds and presented a real-time (on-the-fly) monitoring protocol that guarantees longitudinal privacy of users over timestamps, irrespective of whether their reports are about independent or correlated values. Cheu et al. [118] analyzed the properties of the shuffle model of LDP and proposed distributed differentially private algorithms in the shuffle model; the model is an augmented LDP model with an anonymous channel that randomly permutes a group of user messages. They also demonstrated the power of the shuffled model lies between those of the central and local models. Bale et al. [119] proposed an optimal summation estimation on numerical data in single-message (noninteractive) shuffle model and the method achieved the lower bound in accuracy and privacy amplification bound under the model. Ghazi et al. [120] presented the approximately tight bounds on frequency estimation and the sample complexity in the single-message

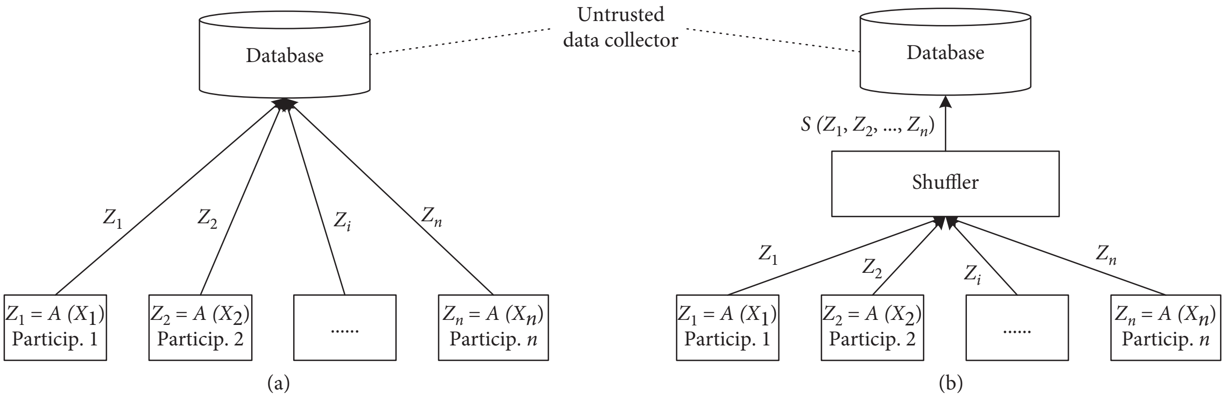


Figure 3: The local and shuttle models for differential privacy (a) The local model (b) The shuttle model.

(noninteractive) shuffle model and proposed several communication-efficient multimessage (multiple-round interactive) statistical estimation (e.g., frequency estimation and range query) protocols with lower bound error.

However, the shuttle model requires that the shuttled server (Shuffler) is trustable and cannot collude with the analyst server (Data collector). If the Shuffler colludes with the analyst server, the Shuffler is redundant and privacy amplification of participants will vanish. Wang et al. [40] introduced the defect of the shuffle model, investigated multiple-party setting differential privacy (MPDP), and analyzed the adversary model and the existing approach. They further proposed novel techniques for obtaining a better tradeoff between privacy and utility than the previous approach. Besides, Feldman et al. [92] focused on the research of private learning problem. Previous differential privacy learning algorithms involves guaranteeing privacy of each step and allows to publish the intermediate results; they demonstrated that contractive iterations can strongly amplify the privacy guarantees under the various variants of differential privacy without depending on publishing intermediate results, and privacy-amplification-by-iteration method can achieve guarantees similar to those obtained using the privacy-amplification-by-sampling technique. However, the iteration method brings into high computation overhead.

6.4. Some Application Fields under LDP. LDP has been adopted in private range query, private multidimensional analytical query, private location data collection, private recommender systems, and other applications. Herein, we introduce the studies on these application fields under LDP.

6.4.1. Private Range Queries under LDP. The previous frequency oracles in the LDP are designed for answering point queries on distribution, namely, frequency estimation. Recently, some studies have focused on the Range queries, which is a fundamental data analysis primitive and can also be used to compute other core statistics and to build prediction models. In this section, we review some methods to support range queries under LDP and its variants. Cormode and Kulkarni et al. [38] introduced and analyzed some methods to implement range queries under the LDP and proposed two approaches for range queries, which are based on hierarchical histograms and Haar wavelet transform. Gu et al. [121] studied the problem of privately answering range queries and achieving frequency estimation with LDP and developed a linear-equations-based mechanism, which satisfies local d-privacy [122] (a general LDP with any distance metric) and achieves optimal utility for co-location query which is a special range query.

6.4.2. Private Multidimensional Analytical (MDA) Queries (Marginal Release). Wang et al. [123] focused on the study of private multidimensional analytical (MDA) queries under the LDP and proposed a weighted frequency oracle as the building block for the MDA queries and proposed several LDP encoder and estimation algorithms to conduct a class of MDA queries with tight error bounds and even scale in high-dimensional situations. Their team [124] further studied the MDA queries under the LDP and proposed an LDP-based middleware solution for differentially private data sharing and analytics as cloud services, called DPSAaS. Note that the above-mentioned studies on marginal release (which is a classic application of histogram estimation) can be seen as a special MDA query (with only count aggregation). There are several studies [62, 63] that focus on the study of marginal release under LDP.

6.4.3. Private Location Data under LDP. There are some studies concentrated on the private location data under LDP. Chen et al. [51] proposed the personalized local differential privacy (PLDP) for private location collection and developed a private count estimation protocol (PCEP) based on SHit protocol to provide good frequency approximation. The core of SHit is the randomized projection matrix that is randomly generated. Sangiamchit et al. [125] investigated on the same problem and developed the PCEP protocol based on the Hadamard Matrix rather than the Randomized Matrix. Recent work conducted by Acharya and Kairouz et al. [19] presented the block-structured LDP model and proposed HRR-based privatization scheme for the practical geo-location data collection application where not all spatial locations are equally sensitive. Kim et al. [126] adopted the LDP for privately collecting the indoor positioning data. They used the RAPPOR mechanism for perturbation in the user-side and a straightforward statistics-based approach as well as an EM-based approach for estimating the density of indoor locations in the server-side. Kim et al. [127] further investigated the problem and proposed a novel workload-aware indoor positioning data aggregation method that can seek an optimal data encoding and perturbation algorithm of LDP to minimize the total estimation error under the given workload. Wang et al. [128] proposed a locally differentially private location data approach for mobile crowdsensing with considering different participants' privacy preferences by choosing two different LDP methods, namely, RAPPOR and GRR. Arcolezi et al. [129] focused on forecasting the number of firemen interventions per location (region) with LDP-based data. In particular, the method adopts LDP to anonymize location data from the users and reconstructs a synthetic dataset from these perturbed data, and then uses extreme gradient boosting (XGBoost), which is a supervised machine learning approach to conduct the prediction. Xiong et al. [130] proposed to apply LDP to private continuous location sharing setting; they introduced a novel variant of LDP to capture the temporal correlations between locations and adopted the GRR mechanism to achieve it for location privacy preservation.

In addition, Geo-Indistinguishability [131] that is used for private location sharing (e.g., location-based services) and its general notion d-privacy [122] also satisfy LDP and can be adopted for locally differentially private location collection. Alvim and Chatzikokolakis et al. [132] proposed a variant of LDP based on d-privacy for distance-sensitive data, e.g., location data. The LDP based on d-privacy can improve the tradeoff between privacy and utility compared to standard LDP approaches.

6.4.4. Private Recommender Systems with LDP. Recent studies have focused on the construction of private recommender systems with LDP. Shen and Hin [133] first investigated the personalized recommendation under the LDP model and proposed a novel relaxed admissible mechanism that achieves the LDP to inject the flexible-instance-based noises for preserving individual's items. They further developed a personalized recommendation system framework called EpicRec [134] enabling the locally differentially private data perturbation. However, the method cannot be suitable for preserving the category attribute preference data of individuals. Hua et al. [135] proposed a differentially private matrix factorization mechanism based on gradient perturbation to protect any individual's ratings or profiles under the untrusted recommender. However, the proposed mechanism is only suitable for protecting individual ratings. Recently, Shin et al. [136] developed a novel matrix factorization algorithm under LDP to enhance privacy by guaranteeing individual privacy and completely protecting items and ratings. Similarly, Asada et al. [137] applied the locally differentially private matrix factorization method proposed by Shin [136] to location privacy preference recommendation. In addition, Zhou et al. [138] proposed a randomized-response-based private recommendation system with high performance and less time and space while locally differentially privacy-preserving individuals' items and ratings.

7. Conclusions and Open Challenges

In this paper, we have conducted a comprehensive and systematic survey on the latest developments in the field of LDP. We have given an overview of the fundamental knowledge and framework of LDP. Then we have introduced the mainstream privatization mechanisms for basic statistical estimation that includes frequency estimation, mean estimation, and distribution estimation under LDP. Further, we have presented the current research circumstances on LDP including the private statistical learning/inferencing and statistical data analysis under LDP, privacy amplification techniques for LDP, and some application fields under LDP. The LDP is a relatively new research field. Although in recent years both academia and industry have made a great effort on the explorations and applications of LDP, there are still inevitably many challenges in the research field of LDP. Some potential open research problems and directions are listed below:

(1) Hybrid model and Shuttle model: Since LDP is a stringent privacy preservation technique at the cost of higher sample complexity and lower accuracy compared to the CDP, a recent study [139] proposed a novel hybrid model that combines the upsides of the LDP model and CDP model to achieve the significant improvement of accuracy in the application of heavy hitter identification. This motivates us that it is an open challenge for extending the hybrid model to other applications. Shuttle model [85] is a novel augmented LDP model for guaranteeing the data and identity privacy of participants. However, the shuttle model was proposed soon and is in the early stages of research; it is an open challenge to apply the shuttle model to investigate the complexity of various statistical and learning tasks and extend to the shuttle model to some stronger privacy-preserving applications.

(2) Multiple round interactivities: The bulk of current LDP protocols require the participants (users) to conduct a fixed protocol over their data and transfer their perturbed data for data analysis (e.g., Aggregation, Learning, and Mining). The protocols for ensuring generalization and statistical validity do not account for the data dependence. Recent studies [32–34] investigate the challenges of adaptive data reuse by answering multiple queries about these data where the aggregator or analyst launches new queries depending arbitrarily on responses to the previous queries, and such protocol is regarded as the multiple rounds of the interactive protocol. The approach has been adopted for heavy hitter identification [67], synthetic graph modeling [78], and learning model

construction [10, 44]. It is open to understand the characteristics of multiple round interactive settings, e.g., the property of compositionality, compared to the noninteractive (single-round interactive) setting. Besides, most theoretical researches on locally private learning focus on the noninteractive (single round) and sequentially interactive protocol. Some recent studies [32, 33] discovered the exponential gap of sample complexity between fully and sequential interactive protocols, which will lead to the problem of the power of locally private learning in the fully interactive setting. It is open to understand the power of multiple round interactivities and the problem of round complexity.

(3) Private Learning Problem. Most studies of locally private learning focused on classical statistical learning, e.g., Empirical risk minimization problem and Hypothesis testing. We highlight the open challenges of studying the LDP empirical risk minimization with nonconvex loss functions and reducing the computation and communication overhead of locally differentially private training of nonconvex models in statistical learning. Locally private federated learning is an active and ongoing research field. The locally private learning requires the model parameter updates computed at every round is private to any parties, e.g., server and all untrusted third parties. Since LDP is too rigorous for practical federal learning applications, especially, LDP with small privacy parameter will extremely affect the accuracy of model-fitting. Hence, how to balance between privacy and model accuracy in private federal learning is extremely challenging. In addition, the model parameter update in private federal learning requires multiple rounds, the number of updates is critical for privacy budget allocation of each round update. Therefore, developing low-update locally private federated learning approaches is also an interesting challenge and direction.

(4) Privacy amplification. LDP model where the server collecting private data from participants can identify any specific participant by retracing their input data can only guarantee the protection of data privacy, but not the protection of identity privacy. Some privacy amplification techniques for LDP, e.g., *k*-anonymity, shuttling, and sampling-iteration can provide the preservation of identity privacy while guaranteeing the data privacy of participants. It is still open how to develop the privacy amplification for the LDP model and how to quantify the degree of privacy amplification under the LDP.

(5) Theoretical foundations. Most studies on LDP have focused on theoretical research issues [30, 32]; however, it is still an open challenge for some theoretical questions: what are the lower bounds on the sample complexity, communication complexity, and even round complexity in the light of privacy parameter for one statistical estimation or learning problem; which encoding technique can achieve the tradeoff between the communication overhead, accuracy guarantee, and privacy guarantee; can the approximate LDP obtain any advantage, compared to the pure LDP definition. In addition, how to quantify and evaluate the theoretical metrics method is also a direction for further research.

(6) Private multidimensional categorical data. Currently, most studies of collecting multidimensional data with LDP mainly focused on the numerical data while providing the proofs. However, there are few solutions [10, 44, 66] that have been proposed to handle multidimensional categorical attributes with LDP via dimensionality reduction. Two previous studies of [10, 44] adopted the sampling technique for dimensionality reduction to reduce the communication cost in multi-dimensional data, but this method inevitably brings about low utility of data due to the usage of sampling technique. Hence, it is still open how to tradeoff between the communication cost and data utility. The work of [66] proposed attributes-splitting method based on the correlation of attributes for dimensionality reduction. Its key is to identify correlated attributes and splitting attributes into low-dimensional attribute clusters. The method can be adopted for pruning the domain of composite attributes. However, for this method, it is also still open how to measure the correlation between attributes. Motivated by the above studies for multidimensional data with LDP, designing efficient solutions for private multidimensional categorical data collection is also an interesting challenge and direction.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Major Technical Innovation Project of Hubei (Grant no. 2018AAA046), National Natural Science Foundation of China (Grant no. 61872431), and Applied Basic Research Project of Wuhan (Grant no. 2017060201010162).

References

[1] C. Dwork, “Differential privacy,” in *Proceedings of the 33rd International Conference on Automata, Languages and Programming (ICALP)*, pp. 1–12, Venice, Italy, July 2006.

[2] L. Sweeney, “*k*-ANONYMITY: a model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-diversity: privacy beyond *k*-anonymity,” in *Proceedings of the 22nd International*

Conference on Data Engineering (ICDE), p. 24, Atlanta, GA, USA, April 2006.

[4] N. Li, T. Li, and S. Venkatasubramanian, "*t*-Closeness: privacy beyond *k*-Anonymity and *l*-diversity," in *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering (ICDE)*, pp. 106–115, Istanbul, Turkey, April 2007.

[5] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately," *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 429–438, Berkeley, CA, USA, October 2013.

[7] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1054–1067, Scottsdale, AZ, USA, November 2014.

[8] Differential Privacy Team Apple, "Learning with privacy at scale," *Apple Machine Learning, Journal*, vol. 1, no. 8, pp. 1–25, 2017.

[9] B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting telemetry data privately," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 3572–3581, Long Beach, CA, USA, December 2017.

[10] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin, "Collecting and analyzing data from smart device users with local differential privacy," 2016, https://arxiv.org/abs/1606.05053.

[11] Q. Ye, X. Meng, M. Zhu, and Z. Huo, "Survey on local differential privacy," *Journal of Software*, vol. 29, no. 7, pp. 1981–2005, 2018.

[12] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: local differential privacy in practice," in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, pp. 1655–1658, Houston, TX, USA, June 2018.

[13] B. Bebensee, "Local differential privacy: a tutorial," 2019, https://arxiv.org/abs/1907.11908.

[14] N. Li and Q. Ye, "Mobile data collection and analysis with local differential privacy," in *Proceedings of the 2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pp. 4–7, Hong Kong, June 2019.

[15] P. Zhao, G. Zhang, S. Wan, G. Liu, and T. Umer, "A survey of local differential privacy for securing internet of vehicles," *The Journal of Supercomputing*, pp. 1–22, 2019.

[16] B. Jiang, M. Li, and R. Tandon, "Context-aware Data Aggregation with Localized Information Privacy," in *Proceedings of the 2018 Communications and Network Security (CNS)*, pp. 1–9, Beijing, China, June 2018.

[17] V. Rastogi, M. Hay, G. Miklau, and D. Suciu, "Relationship privacy: output perturbation for queries with joins," in *Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 107–116, Washington, DC, USA, May 2009.

[18] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5018–5029, 2016.

[19] J. Acharya, K. Bonawitz, P. Kairouz, D. Ramage, and Z. Sun, "Context-aware local differential privacy," 2019, https://arxiv.org/abs/1911.00038.

[20] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Theory of Cryptography Conference*, pp. 265–284, New York, NY, USA, March 2006.

[21] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 94–103, Providence, RI, USA, October 2007.

[22] 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.

[23] A. D. Sarwate and L. Sankar, "A rate-disortion perspective on local differential privacy," in *Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 903–908, Monticello, IL, USA, 2014.

[24] K. Kalantari, L. Sankar, and A. D. Sarwate, "Robust privacy-utility tradeoffs under differential privacy and hamming distortion," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2816–2830, 2018.

[25] S. Xiong, A. D. Sarwate, and N. B. Mandayam, "Randomized requantization with local differential privacy," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2189–2193, Shanghai, China, March 2016.

[26] M. Lopuhaä-Zwakenberg, B. Škorić, and N. Li, "Information-theoretic metrics for local differential privacy protocols," 2019, https://arxiv.org/abs/1910.07826.

[27] I. Wagner and D. Eckhoff, "Technical privacy metrics," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–38, Jul. 2018.

[28] M. Lopuhaä-Zwakenberg, Z. Li, B. Škorić, and N. Li, "Four accuracy bounds and one estimator for frequency estimation under local differential privacy," 2019, https://arxiv.org/abs/1911.10499.

[29] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 2879–2887, Montreal, Quebec, Canada, December 2014.

[30] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," 2013, https://arxiv.org/abs/1302.3203.

[31] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–201, 2018.

[32] M. Joseph, J. Mao, and A. Roth, "Exponential separations in local differential privacy," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 515–527, Washington, DC, USA, January 2020.

[33] M. Joseph, J. Mao, S. Neel, and A. Roth, "The role of interactivity in local differential privacy," in *Proceedings of the IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105, Baltimore, MD, USA, November 2019.

[34] J. Duchi and R. Rogers, "Lower bounds for locally private estimation via communication complexity," in *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 1161–1191, Phoenix, AZ, USA, 2019.

[35] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proceedings of the*

33rd International Conference on Machine Learning (ICML), pp. 2436–2444, New York, NY, USA, June 2016.

[36] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proceedings of the 26th USENIX Security Symposium*, pp. 729–745, Vancouver, BC, Canada, August 2017.

[37] R. Bassily and A. Smith, "Local, private, efficient protocols for Succinct histograms," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 127–135, Portland, OR, USA, June 2015.

[38] G. Cormode, T. Kulkarni, and D. Srivastava, "Answering range queries under local differential privacy," *Proceedings of the VLDB Endowment*, vol. 12, no. 10, pp. 1126–1138, 2019.

[39] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: estimating distributions privately, efficiently, and with little communication," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1120–1129, Naha, Japan, April 2019.

[40] T. Wang, B. Ding, M. Xu et al., "MURS: practical and robust privacy amplification with multi-party differential privacy," 2019, https://arxiv.org/abs/1908.11515v2.

[41] S. Wang, L. Huang, P. Wang et al., "Mutual information optimally local private discrete distribution estimation," 2016, https://arxiv.org/abs/1607.08025.

[42] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5662–5676, Aug. 2018.

[43] Y. Sei and A. Ohsuga, "Differential private data collection and analysis based on randomized multiple Dummies for untrusted mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 926–939, 2017.

[44] N. Wang, X. Xiao, Y. Yang et al., "Collecting and Analyzing Multidimensional Data with Local Differential Privacy," in *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 638–649, Macao, China, 2019.

[45] J. Jia and N. Z. Gong, "Calibrate: frequency estimation and heavy hitter identification with local differential privacy via incorporating prior knowledge," in *Proceedings of the IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2008–2016, Paris, France, April 2019.

[46] T. Murakami and Y. Kawamoto, "Utility-optimized local differential privacy mechanisms for distribution estimation," in *Proceedings of the 28th USENIX Security Symposium*, pp. 1877–1894, Santa Clara, CA, USA, August 2019.

[47] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy," *Information Sciences*, vol. 250, pp. 200–214, 2013.

[48] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1176–1184, 2015.

[49] T. Wang, J. Zhao, X. Yang, and X. Ren, "Locally differentially private data collection and analysis," 2019, https://arxiv.org/abs/1906.01777.

[50] M. Akter and T. Hashem, "Computing aggregates over numeric data with personalized local differential privacy," in *Proceedings of the Australasian Conference on Information Security and Privacy*, pp. 249–260, Auckland, New Zealand, July 2017.

[51] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in *Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 289–300, Helsinki, Finland, May 2016.

[52] Z. Li, T. Wang, M. Lopuhaä-Zwakenberg, N. Li, and B. Škoric, "Estimating numerical distributions under local differential privacy," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 621–635, Portland, OR, USA, June 2020.

[53] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77, San Jose, CA, USA, May 2017.

[54] M. Gaboardi, R. Rogers, and O. Sheffet, "Locally private mean estimation: Z-test and tight confidence intervals," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2545–2554, Naha, Okinawa, Japan, April 2019.

[55] M. Joseph, J. Kulkarni, J. Mao, and Z. S. Wu, "Locally private Gaussian estimation," 2018, https://arxiv.org/abs/1811.08382.

[56] N. Holohan, D. J. Leith, and O. Mason, "Optimal differentially private mechanisms for randomised response," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2726–2735, 2017.

[57] G. Fanti, V. Pihur, and Ú. Erlingsson, "Building a RAPPOR with the unknown: privacy-preserving learning of associations and data dictionaries," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 3, pp. 41–61, 2016.

[58] Y. Ye, M. Zhang, D. Feng, H. Li, and J. Chi, "Multiple privacy regimes mechanism for local differential privacy," in *Proceedings of the International Conference on Database Systems for Advanced Applications*, pp. 247–263, Chiang Mai, Thailand, April 2019.

[59] T. Murakami, H. Hino, and J. Sakuma, "Toward distribution estimation under local differential privacy with small samples," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 84–104, 2018.

[60] M. E. Gursoy, A. Tamersoy, S. Truex, W. Wei, and L. Liu, "Secure and utility-aware data collection with condensed local differential privacy," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2019.

[61] R. Bassily, "Linear queries estimation with local differential privacy," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 721–729, Naha, Okinawa, Japan, April 2019.

[62] G. Cormode, T. Kulkarni, and D. Srivastava, "Marginal release under local differential privacy," in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, pp. 131–146, Houston, TX, USA, June 2018.

[63] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "CALM: Consistent adaptive local marginal for marginal release under local differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 212–229, Toronto, Canada, October 2018.

[64] F. Peng, S. Tang, B. Zhao, and Y. Liu, "A privacy-preserving data aggregation of mobile crowdsensing based on local differential privacy," in *Proceedings of the ACM Turing Celebration Conference-China*, pp. 1–5, Chengdu, China, May 2019.

[65] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, and J. McCann, "High-dimensional crowdsourced data distribution estimation with local privacy," in *Proceedings of the 2016 IEEE International Conference on Computer and Information Technology*, pp. 226–233, Nadi, Fiji, 2016.

[66] X. Ren, C.-M. Yu, W. Yu et al., "LoPub: high-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.

[67] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 192–203, Vienna, Austria, October 2016.

[68] T. Wang, N. Li, and S. Jha, "Locally differentially private frequent itemset mining," in *Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP)*, pp. 127–143, San Francisco, CA, USA, May 2018.

[69] S. Wang, L. Huang, Y. Nie, P. Wang, W. Yang, and H. Xu, "PrivSet: set-valued data analyses with locale differential privacy," in *Proceedings of the IEEE INFOCOM 2018 --IEEE Conference on Computer Communications*, pp. 1088–1096, Honolulu, HI, USA, April 2018.

[70] Q. Ye, H. Hu, X. Meng, and H. Zheng, "PrivKV: key-value data collection with local differential privacy," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pp. 317–331, San Francisco, CA, USA, May 2019.

[71] J. Yang, X. Cheng, S. Su, R. Chen, Q. Ren, and Y. Liu, "Collecting preference rankings under local differential privacy," in *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1598–1601, Macao, April 2019.

[72] Z. Yan, G. Li, and J. Liu, "Private rank aggregation under local differential privacy," *International Journal of Intelligent Systems*, vol. 35, no. 35, pp. 1492–1519, 2020.

[73] S. Wang, Y. Nie, P. Wang, H. Xu, W. Yang, and L. Huang, "Local private ordinal data distribution estimation," in *Proceedings of the IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, May 2017.

[74] S. Wang, L. Huang, Y. Nie et al., "Local differential private data aggregation for discrete distribution estimation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 9, pp. 2046–2059, 2019.

[75] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan et al., "Learning new words," US Patent 9594741B1, 2017.

[76] S. Kim, H. Shin, C. Baek, S. Kim, and J. Shin, "Learning new words from keystroke data with local differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 479–491, 2020.

[77] N. Wang, X. Xiao, Y. Yang et al., "PrivTrie: effective frequent term discovery under local differential privacy," in *Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 821–832, Paris, France, April 2018.

[78] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren, "Generating synthetic decentralized social graphs with local differential privacy," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 425–438, Dallas, TX, USA, October 2017.

[79] Y. Zhang, J. Wei, X. Zhang, X. Hu, and W. Liu, "A two-phase algorithm for generating synthetic graph under local differential privacy," in *Proceedings of the 8th International Conference on Communication and Network Security*, pp. 84–89, Qingdao, China, November 2018.

[80] T. Gao, F. Li, Y. Chen, and X. Zou, "Local differential privately anonymizing online social networks under HRG-based model," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 1009–1020, 2018.

[81] H. Sun, X. Xiao, I. Khalil et al., "Analyzing subgraph statistics from extended local views with decentralized differential privacy," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 703–717, London UK, November 2019.

[82] C. Wei, S. Ji, C. Liu, W. Chen, and T. Wang, "AsgLDP: collecting and generating decentralized attributed graphs with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3239–3254, 2020.

[83] D. Zhao, H. Chen, S. Zhao, X. Zhang, C. Li, and R. Liu, "Local differential privacy with K-anonymous for frequency estimation," in *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*, pp. 5819–5828, Los Angeles, CA, USA, December 2019.

[84] M. Joseph, A. Roth, J. Ullman, and B. Waggoner, "Local differential privacy for evolving data," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 2375–2384, Montreal, Canada, December 2018.

[85] A. Bittau, Ú. Erlingsson, P. Maniatis et al., "PROCHLO: strong privacy for analytics in the crowd," in *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 441–459, Shanghai, China, October 2017.

[86] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: from local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479, San Diego, CA, USA, January 2019.

[87] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," *Journal of the ACM*, vol. 61, no. 6, pp. 1–57, 2014.

[88] K. Zheng, W. Mou, and L. Wang, "Collect at once, use effectively: making non-interactive locally private learning possible," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 4130–4139, Sydney, Australia, August 2017.

[89] D. Wang, M. Gaboardi, and J. Xu, "Empirical risk minimization in non-interactive local differential privacy revisited," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 965–974, Montreal, Canada, December 2018.

[90] D. Wang, A. Smith, and J. Xu, "Noninteractive locally private learning of linear models via polynomial approximations," in *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, vol. 98, pp. 898–903, Chicago, IL, USA, March 2019.

[91] D. Wang and J. Xu, "On sparse linear regression in the local differential privacy model," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6628–6637, Long Beach, CA, USA, June 2019.

[92] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," in *Proceedings of the 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532, Paris, France, October 2018.

[93] D. Van Der Hoeven, "User-specified local differential privacy in unconstrained adaptive online learning," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 14103–14112, Vancouver, BC, Canada, May 2019.

[94] M. Gaboardi and R. Rogers, "Local private hypothesis testing: chi-square tests," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 1626–1635, Stockholm, Sweden, July 2018.

[95] O. Sheffet, "Locally private hypothesis testing," in *Proceedings of the 35th International Conference on Machine Learning*, pp. 4605–4614, Stockholm, Sweden, July 2018.

[96] J. Acharya, C. L. Canonne, C. Freitag, and H. Tyagi, "Test without trust: optimal locally private distribution testing," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2067–2076, Naha, Japan, April 2019.

[97] C. L. Canonne, G. Kamath, A. McMillan, A. Smith, and J. Ullman, "The structure of optimal private tests for simple hypotheses," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 310–321, Phoenix, AZ, USA, June 2019.

[98] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: a client level perspective," 2017, https://arxiv.org/abs/1712.07557.

[99] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada, April 2018.

[100] J. Li, M. Khodak, S. Caldas, and A. Talwalkar, "Differentially private meta-learning," 2019, https://arxiv.org/abs/1909.05830.

[101] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018, https://arxiv.org/abs/1812.00984.

[102] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "Local differential privacy for deep learning," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5827–5842, 2020.

[103] C. Xu, J. Ren, L. She et al., "EdgeSanitizer: locally differentially private deep inference at the edge for mobile data analytics," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5140–5151, 2019.

[104] D. Wang and J. Xu, "Differentially private high dimensional sparse covariance matrix estimation," 2019, https://arxiv.org/abs/1901.06413.

[105] D. Wang and J. Xu, "Lower bound of locally differentially private sparse covariance matrix estimation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4788–4794, Macao, August 2019.

[106] M. F. Balcan, S. S. Du, Y. Wang, and A. W. Yu, "An improved gap-dependency analysis of the noisy power method," in *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 284–309, New York, NY, USA, 2016.

[107] J. Ge, Z. Wang, M. Wang, and H. Liu, "Minimax-optimal privacy-preserving sparse PCA in distributed systems,," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1589–1598, Canary Islands, Spain, April 2018.

[108] D. Wang and J. Xu, "Principal component analysis in the local differential privacy model," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4795–4801, Macao, August 2019.

[109] K. Nissim and U. Stemmer, "Clustering algorithms for the centralized and local models," in *Proceedings of the Algorithmic Learning Theory (ALT)*, pp. 619–653, Lanzarote, Spain, April 2018.

[110] R. Bassily, K. Nissim, U. Stemmer, and A. Thakurta, "Practical locally private heavy hitters," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 2288–2296, Long Beach, CA, USA, 2017.

[111] H. Kaplan and U. Stemmer, "Differentially private *K*-means with constant multiplicative error," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 5436–5446, Montreal, Canada, December 2018.

[112] J. Hsu, S. Khanna, and A. Roth, "Distributed private heavy hitters," in *Proceedings of the 39th International Colloquium Conference on Automata, Languages, and Programming*, pp. 461–472, Warwick, UK, July 2012.

[113] N. Mishra and M. Sandler, "Privacy via pseudorandom sketches," in *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 143–152, San Diego, CA, USA, March 2006.

[114] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 85–94, Long Beach, CA, USA, June 2019.

[115] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," in *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pp. 435–447, Houston, TX, USA, June 2018.

[116] T. Wang, N. Li, and S. Jha, "Locally differentially private heavy hitter identification," *IEEE Transactions on Dependable and Secure Computing*, vol. 5971, pp. 1–12, 2019.

[117] X. Zhang, L. Huang, P. Fang, S. Wang, Z. Zhenyu, and H. Xu, "Differentially private frequent itemset mining from smart devices in local setting," in *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications*, pp. 433–444, Guilin, China, June 2017.

[118] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," in *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 375–403, Darmstadt, Germany, May 2019.

[119] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Proceedings of the Annual International Cryptology Conference*, pp. 638–667, Santa Barbara, CA, USA, August 2019.

[120] B. Ghazi, N. Golowich, R. Kumar, R. Pagh, and A. Velingker, "On the power of multiple anonymous messages," 2019, https://arxiv.org/abs/1908.11358.

[121] X. Gu, M. Li, Y. Cao, and L. Xiong, "Supporting both range queries and frequency estimation with local differential privacy," in *Proceedings of the 2019 IEEE Conference on Communications and Network Security*, pp. 124–132, Washington, DC, USA, 2019.

[122] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Proceedings of the 13th International Symposium on Privacy Enhancing Technologies (PETS)*, pp. 82–102, Bloomington, IN, USA, July 2013.

[123] T. Wang, B. Ding, J. Zhou et al., "Answering multi-dimensional analytical queries under local differential privacy," in *Proceedings of the 2019 International Conference on Management of Data (SIGMOD)*, pp. 159–176, Amsterdam Netherlands, June 2019.

[124] M. Xu, T. Wang, B. Ding et al., "Multi-dimensional data sharing and analytics as services under local differential privacy," *Proceedings of the VLDB Endowment*, vol. 12, no. 12, pp. 1862–1865, 2019.

[125] P. Sangiamchit and J. Fakcharoenphol, "Practical differential privacy for location data aggregation using a Hadamard matrix," in *Proceedings of the 2019 16th International Joint*

Conference on Computer Science and Software Engineering (JCSSE), pp. 79–84, Pataya, Thailand, 2019.

[126] J. W. Kim, D. H. Kim, and B. Jang, "Application of local differential privacy to collection of indoor positioning data," *IEEE Access*, vol. 6, pp. 4276–4286, 2018.

[127] J. W. Kim and B. Jang, "Workload-aware indoor positioning data collection via local differential privacy," *IEEE Communications Letters*, vol. 23, no. 8, pp. 1352–1356, 2019.

[128] J. Wang, Y. Wang, G. Zhao, and Z. Zhao, "Location protection method for mobile crowd sensing based on local differential privacy preference," *Peer-to-Peer Networking and Applications*, vol. 12, no. 5, pp. 1097–1109, 2019.

[129] H. H. Arcolezi, "Forecasting the number of firefighter interventions per region with local-differential-privacy-based data," *Computers & Security*, vol. 96, pp. 1–12, 2020.

[130] X. Xiong, S. Liu, D. Li, J. Wang, and X. Niu, "Locally differentially private continuous location sharing with randomized response," *International Journal of Distributed Sensor Networks*, vol. 15, no. 8, Article ID 155014771987037, 2019.

[131] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pp. 901–914, Berlin, Germany, November 2013.

[132] M. Alvim, K. Chatzikokolakis, C. Palamidessi, and A. Pazii, "Local Differential Privacy on Metric Spaces: Optimizing the Trade-Off with Utility," in *Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 262–267, Oxford, UK, July 2018.

[133] Y. Shen and H. Jin, "Privacy-preserving personalized recommendation: an instance-based approach via differential privacy," in *Proceedings of the 2014 IEEE International Conference on Data Mining*, pp. 540–549, Shenzhen, China, 2014.

[134] Y. Shen and H. Jin, "EpicRec: towards practical differentially private framework for personalized recommendation," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 180–191, Vienna, Austria, October 2016.

[135] J. Hua, C. Xia, and S. Zhong, "Differentially private matrix factorization," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1763–1770, Buenos Aires, Argentina, July 2015.

[136] H. Shin, S. Kim, J. Shin, and X. Xiao, "Privacy enhanced matrix factorization for recommendation with local differential privacy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1770–1782, 2018.

[137] M. Asada, M. Yoshikawa, and Y. Cao, "'When and where do you want to hide?'–recommendation of location privacy preferences with local differential privacy," in *Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 164–176, Charleston, SC, USA, July 2019.

[138] H. Zhou, G. Yang, Y. Xu, and W. Wang, "Effective matrix factorization for recommendation with local differential privacy," in *Proceedings of the International Conference on Science of Cyber Security*, pp. 235–249, Nanjing, China, August 2019.

[139] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits, "Blender: enabling local search with a hybrid differential privacy model," in *Proceedings of the 26th USENIX Security Symposium*, pp. 747–764, Vancouver, BC, Canada, August 2017.