Supporting both Range Queries and Frequency Estimation with Local Differential Privacy

Xiaolan Gu*, Ming Li*, Yang Cao† and Li Xiong‡
*Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA
†Department of Social Informatics, Kyoto University, Kyoto, Japan
‡Department of Computer Science, Emory University, Atlanta, GA, USA
Email: *{xiaolang, lim}@email.arizona.edu, †yang@i.kyoto-u.ac.jp, ‡lxiong@emory.edu

***Abstract*—Local Differential Privacy (LDP) provides provable privacy protection for data collection without the assumption of the trusted data server. Existing mechanisms that satisfy LDP or its variants either only consider aggregate queries from a group of users (e.g., frequency estimation) or individual queries for a single user (e.g., range queries). However, in complex real-world analytics applications, it is desirable to support both types of queries at the same time.**

In this paper, we tackle the challenge of privately answering range queries and providing frequency estimation at the same time with high utility. We develop a data perturbation mechanism, which is proved to satisfy local d-privacy (a generalized version of LDP with distance metric) and have optimal utility for the co-location query (a specific type of range query). Then, we utilize an inversion approach for frequency estimation using the perturbed data. We analyze the theoretical Mean Square Error (MSE) of this estimation method and show the relationship to another existing estimation method under LDP. The results on both synthetic and real-world location datasets validate the correctness of our theoretical analysis and show that the proposed mechanism has better utility for both range queries and frequency estimation than the state-of-the-art mechanisms.

I. Introduction

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

This work was partly supported by NSF grants CNS-1731164 and No. 1618932, the AFOSR DDDAS program under grant FA9550-121-0240, and Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (S) No. 17H06099 and (A) No. 18H04093.

In the LDP setting, the most existing mechanisms [3], [4], [6], [7] are studied in the context of aggregate queries (e.g., frequency estimation) and the utility requires the aggregation of the perturbed values from a large group of users, while the individual perturbed value may not provide much utility. In order to provide reasonable utility of the perturbed value for an individual user, variants of LDP have been studied. For example, in the Location-Based Services (LBS) setting where users submit their locations to a service provider for range queries (e.g., finding the nearest restaurants), the goal is to accurately answer the range query using the perturbed location. In order to provide better utility in such cases, Chatzikokolakis [8] et al. presented a generalized notion, termed d-privacy, which scales the privacy in DP with the distance between the input pairs. When Euclidean distance is considered in the local setting, the d-privacy is called geo-indistinguishability [9], which is often used in the LBS setting. The intuition is that we require a stronger indistinguishability between two locations when they are close to each other, i.e. the indistinguishabilty is scaled by the distance between two input locations.

However, existing studies on LDP or its variants only focus on one type of queries, i.e. frequency estimation or range queries, but not simultaneously. Our primary viewpoint in this paper is that for certain applications, it is desirable to support both types of queries with high utility. For example, in the LBS setting, suppose a location service provider is offering services such as POI (points of interest) search. In order to protect users' location privacy, users can submit their perturbed location to the service provider. The service provider then uses the perturbed location to answer individual ranges queries to support POI search. At the same time, the service provider also wishes to aggregate the perturbed data from all the users to analyze trends, e.g., number of users at a certain location (i.e., frequency estimation). Another example is the study of social relationship analysis from location data [10], [11], which is more accurate than from self-report survey data. The co-location events of two users and the location entropy (i.e., the expected number of users visiting a location) are both needed to compute the social strength between them. The former indicates whether they frequently meet with each other; the latter indicates the attribute of the meeting places such as a public hotspot or a private place. Since only users

who frequently meet in private places can be considered to have strong social ties, both types of queries must be answered with high accuracy to get the good utility of social relationship analysis.

In this paper, we focus on the d-privacy notion in the local setting, i.e., local d-privacy (a generalized version of LDP), and two specific types of queries: range queries that involve individual information and frequency estimation that involves aggregate information. The former asks the value range (such as location range) of a specific user, the latter asks how many users have the value of interest or are visiting a location. To the best of our knowledge, existing mechanisms that satisfy LDP or its variants only deal with one of these queries. For example, based on a surveying technique termed Randomized Response [12], perturbation mechanisms [3], [4], [6], [7] are developed for frequency estimation under LDP and have good performance. These mechanisms do not consider individual queries such as range queries in their design goals, which may yield poor utility for the latter. On the other hand, some noise-adding mechanisms have been shown to provide high utility for range queries in the local setting. For example, Andrés et al. [9] proposed the Planar Laplace mechanism (satisfying geo-indistinguishability) which adds two-dimensional Laplace noise to a user's location, where a higher probability is associated with an output location that is closer to the original location. However, this mechanism is only applicable to the Euclidean distance metric and is designed for range queries not for frequency estimation, so it may not be suitable for the latter. Thus, it still remains an open question to design a privacy-preserving mechanism that can well support both types of queries.

Intuitively, since randomized response mechanisms have good performance for frequency estimation while Planar Laplace Mechanism yields better utility for range queries, we can combine these two types of mechanisms to get the best of both worlds. The basic idea is to assign different perturbation probabilities for different inputs and outputs during the randomized response in a way related to the distance. However, there are several challenges to design such a mechanism: (1) The mechanism design can be formulated as an optimization problem with the goal of maximizing both utilities while satisfying local d-privacy with any distance metric, but solving this problem requires a high computational cost due to a large number of privacy constraints. (2) The utility function (or expected error) of various types of queries (e.g., frequency estimation) may not be easily expressed in closed form, and even if possible, the function can be non-linear and non-convex which makes the optimization problem hard to solve. (3) Planar Laplace Mechanism can only handle Euclidean distance metric thus has limited applications. Therefore, it is necessary to design a mechanism satisfying local d-privacy that can handle any distance metric with affordable computational cost.

The main contributions of this paper are summarized below:

(1) We apply the notion of local d-privacy to privacy protection to support both individual range queries for each single user and frequency estimation from multiple users. We present a utility-privacy optimization framework with the goal of maximizing the utility while satisfying the local d-privacy constraints.

(2) Considering a large number of constraints and the complex objective function, we develop a new mechanism, which solves the linear equations of perturbation probabilities, to obtain the solution rather than solving the optimization problem directly. We prove that this mechanism satisfies local d-privacy with any distance metric and can obtain the optimal result for co-location query (a specific type of range query).

(3) We utilize the inversion approach to implement frequency estimation of our mechanism under local d-privacy and analyze the theoretical Mean Square Error (MSE). We also show the relationship to an existing estimation method in the LDP setting.

(4) We validate the effectiveness of our mechanism and the correctness of the theoretical analysis by experiments on both synthetic and real-world location datasets. It turns out that the proposed mechanism outperforms the existing ones for both range queries and frequency estimation. Moreover, empirical results show that domain size has relatively little impact on the utility of mechanisms that satisfy local d-privacy, while the MSE for frequency estimation of mechanisms satisfying LDP is proportional to the domain size.

II. Related Work

There are several mechanisms designed for answering individual queries (e.g., range queries) with LDP guarantee. Exponential Mechanism [13] provides a differentially private selection from a discrete set of candidate outputs. It relies on a score function that assigns a valued score to a pair of input-output, where higher scores indicate more desirable outputs related to the input. Planar Laplace Mechanism [9] is an extended version of the Laplace Mechanism [2] (a mechanism for traditional DP) in the planar scenario. It satisfies geo-indistinguishability but only uses the Euclidean distance metric. Chatzikokolakis et al. [14] studied the optimal trade-off between utility and privacy under local d-privacy with an arbitrary distance metric, and formulated it as a linear programming (LP) problem, where any distance metric is applicable. Considering solving the LP problem might suffer from a high computational cost due to a large number of privacy constraints, the authors reduced the number of constraints via a graph-based approximation technique, but their result is not optimal.

For answering aggregate queries (e.g., frequency estimation), Erlingsson et al. [4] developed RAPPOR satisfying LDP for Chrome to collect URL click counts. It is based on the ideas of Randomized Response [12], which is a technique for collecting statistics on sensitive queries when a respondent wants to retain confidentiality. In the basic RAPPOR, they adopt unary encoding to obtain better performance of frequency estimation. Wang et al. [6] optimized the parameters of basic RAPPOR by minimizing the variance of frequency estimation. Qin et al. [15] developed a two-phase mechanism LDPMiner to obtaining accurate heavy hitters with LDP.

Some other works studied privacy-preserving techniques in context-aware scenario, which consider prior knowledge to improve the utility. For example, Pingley et al. [16] proposed the context-aware privacy protection system for LBS, and Jiang et al. [17] presented Localized Information Privacy (LIP) for context-aware data aggregation. In this paper, we study the privacy-preserving techniques under LDP and its variants, which provide strong protection that is independent of the adversary's prior knowledge (i.e. context-free scenario).

III. Problem Formulation

A. System Model and Threat Model

Our system model involves one data server and n users $\mathcal{U} = \{u_1, u_2 \cdots, u_n\}$. Each user has one record (discrete data) and perturbs it independently via a random perturbation mechanism $\mathcal{M}$ before uploading it to the server. Then, the server collects users' data and learns some information to answer the queries requested by users or a third party. For the perturbation mechanism $\mathcal{M}$, we assume the input domain and output domain are the same $\mathcal{D} = \{l_1, l_2, \cdots, l_m\}$, where m is the domain size, and element l_i has finite dimension. For example, l_i can be a numerical value for survey collection, or a two-dimensional vector representing two coordinate values of the rectangular coordinate system for locations. With the discrete setting, the integer set $\mathcal{I} = \{1, 2, \cdots, m\}$ can be used to index the elements in domain $\mathcal{D}$. Then, the randomized response mechanism $\mathcal{M}$ can be implemented by a probability matrix $\mathbf{P} \in [0,1]^{m\times m}$, where the element p_{ij} denotes the probability that the input $x = l_i$ (raw data) is perturbed to the output $y = l_j$ (perturbed data). In this paper, we consider two kinds of specific queries: range queries and frequency estimation. The former asks whether the value of a specific user u_t is in a range $\mathcal{R}$, then the server would return a positive or negative response. The latter asks how many users have an interested value l_i, then the server would return a corresponding counting result.

In the threat model, we assume the server is untrusted and each user only trusts herself because the privacy leakage can be caused by either deliberate commercial transactions or hacking activities. Therefore, the adversary is assumed to possess the uploaded (perturbed) data of all users and know the perturbation mechanism adopted by the users. In this paper, we only focus on providing event-level privacy guarantees [9], [14] (protecting each individual record) rather than the user-level such as protecting the whole location trace of a user [18], [19]. The latter will be our future work.

B. Privacy and Utility Definitions

We first review two privacy definitions and then explain why the second one is adopted in this paper.

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

$$\frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} \leqslant e^{\epsilon} \tag{1}$$

where $Range(\mathcal{M})$ is the set of all possible outputs of the mechanism $\mathcal{M}$. The smaller ϵ indicates a stronger privacy protection.

Definition 2 *(Local d-Privacy [8]) For a given $\epsilon \in \mathbb{R}^+$, randomized mechanism $\mathcal{M}$ satisfies local d-privacy if and only if for any input x, x', and any output $y \in Range(\mathcal{M})$*

$$\frac{\Pr(\mathcal{M}(x) = y)}{\Pr(\mathcal{M}(x') = y)} \leqslant e^{\epsilon \cdot d(x,x')} \tag{2}$$

where $d(\cdot,\cdot)$ is a distance metric, which satisfies three properties by definition: $d(x,x) = 0$, $d(x,x') = d(x',x)$ for any x, x', and triangle inequality

$$d(x,x') + d(x,x'') \geqslant d(x',x''), \quad \forall x, x', x'' \tag{3}$$

Local d-privacy is a generalized version of LDP. It introduces the distance metric that scales the privacy with the distance between the input pairs and relaxes the privacy constraint when $d(x,x') > 1$, thus can provide better utility especially for individual range queries. Selection of the distance metric depends on the practical scenario and data format. For example, Euclidean distance is suitable for location data in location-based systems, while Hamming distance is often used for databases. In this paper, we consider local d-privacy with any distance metric so it is generalizable to various application scenarios.

Since the probability matrix $\mathbf{P} = [p_{ik}]_{m\times m}$ determines the perturbation mechanism $\mathcal{M}$, and the input (output) domain $\mathcal{D}$ is indexed by integer set $\mathcal{I}$, we can rewrite the privacy constraint of local d-privacy as

$$p_{ik}/p_{jk} \leqslant e^{\epsilon d_{ij}} \;\Rightarrow\; p_{ik} - e^{\epsilon d_{ij}} \cdot p_{jk} \leqslant 0 \quad (\forall i,j,k) \tag{4}$$

where the element of probability matrix $\mathbf{P}$ is defined by $p_{ik} = \Pr(y = k | x = i)$, and $d_{ij} = d(x = i, x' = j)$ $(i,j \in \mathcal{I})$ is the distance between i and j.

Utility of range query. The accuracy (quantified by false-positive and false-negative) of the server's response to a range query depends on not only the mechanism but also the requested range, which makes the evaluation more complex. Considering both a larger range size of range query and a smaller distance between input and output data would yield a higher accuracy of the response, we utilize the portion of users whose perturbed data is outside a range (with size r) of the raw data to approximate the error rate for the range queries

$$\text{Error}_{\text{range}} = \frac{1}{n}\sum\nolimits_{t=1}^{n}(1 - \mathbb{1}_{\mathcal{R}_{x_t,r}}(y_t)) \tag{5}$$

where x_t, y_t are the raw data and perturbed data for user u_t, $\mathcal{R}_{x_t,r} \triangleq \{k | k \in \mathcal{I}, d(k,x_t) \leqslant r\}$ is the neighboring set of x_t in range size r. Indicator function $\mathbb{1}_{\mathcal{R}_{x_t,r}}(y_t)$ equals to 1 when $y_t \in \mathcal{R}_{x_t,r}$ and equals to 0 otherwise. Then the theoretical Error (expectation) is

$$\text{Error}_{\text{range}} = \frac{1}{n}\sum\nolimits_{t=1}^{n}\left(1 - \sum\nolimits_{y_t \in \mathcal{R}_{x_t,r}} p_{x_t y_t}\right) \tag{6}$$

Particularly, $\text{Error}_{\text{range}} = 1 - \frac{1}{n}\sum_{t=1}^{n} p_{x_t x_t}$ when $r = 0$.

Utility of frequency estimation. The utility of frequency estimation is quantified by the Mean Square Error (MSE) of a frequency estimator, where the MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as

$$\text{MSE}_{\hat{\theta}} = \mathbb{E}_{\hat{\theta}}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2 \tag{7}$$

The estimator is usually designed to be unbiased, i.e., $\text{Bias}(\hat{\theta}, \theta) = 0$, hence the MSE is equivalent to the variance of $\hat{\theta}$. We compute the overall MSE as the summation of the MSEs of frequency estimators at all l_k. More details of frequency estimator and MSE analysis are shown in Section V.

C. Objectives and Challenges

In general, our goal is to minimize the Error of range query and the MSE of frequency estimation at the same time while satisfying local d-privacy with any distance metric. The design of probabilities in the perturbation matrix can be represented by the following optimization problem

$$\begin{aligned} \min_{0 \leqslant p_{ij} \leqslant 1} \quad & f(p) \\ \text{s.t.} \quad & g_{ij}^{(k)}(p) \triangleq p_{ik} - e^{\epsilon d_{ij}} \cdot p_{jk} \leqslant 0 \quad (\forall i, j \in \mathcal{I}) \\ & h_k(p) \triangleq \sum\nolimits_{i=1}^{m} p_{ki} - 1 = 0 \quad (\forall k \in \mathcal{I}) \end{aligned} \tag{8}$$

where $p \in \mathbb{R}^{m \times m}$ is the optimization variable whose elements are p_{ij} $(i, j \in \mathcal{I})$. The inequality constraint function $g_{ij}^{(k)}$ is obtained by (4), where $g_{ij}^{(k)} = 0$ for $i = j$. The equality constraint function h_k is obtained by the property of probability matrix. The objective function in (8) can be defined by any linear combination of utility for range query and frequency estimation, i.e., $f(p) = \alpha \cdot \text{Error}_{\text{range}} + \beta \cdot \text{MSE}_{\text{freq}}$, where $\alpha, \beta \geqslant 0$ are the combination coefficients whose values can be determined according to application requirements.

There are several challenges to solve this problem. (1) The existing mechanisms only consider one type of query; thus they might have bad performance on a different type of query or even unable to handle it. (2) The notion of local d-privacy can provide better utility than LDP due to the relaxed privacy constraints, but designing a mechanism applicable to any distance metric or solving the optimization problem becomes much harder because of the large number of variables and constraints, especially when domain size is large. For example, assume domain size is m, then the optimization problem has m^2 variables, m^3 inequality constraints, m equality constraints, and m^2 lower-bound and upper-bound constraints according to the privacy definition. In [14], the original constraints are transformed into $O(m^2)$ constraints, but the result is not optimal after this transformation. (3) The theoretical MSE of a frequency estimator is often related to the true frequency, which is unknown in practice because each user only uploads the perturbed data instead of the raw data in the local setting. Thus, it is difficult to get the accurate theoretical MSE, then the solution of the optimization problem that minimizes the theoretical MSE may not provide optimal performance in practice.

IV. Proposed Data Perturbation Mechanism

Considering the complex objective function and a large number of constraints in (8), we adopt another way with an affordable computational cost to obtain a good utility. First, we present a linear equations based mechanism, which satisfies local d-privacy with any distance metric, with the goal of optimizing the utility for range queries. Then, the optimality of this mechanism for co-location query (a specific range query) is proved via optimization theorem. We will extend this mechanism for frequency estimation and analyze its utility in the next section.

A. Linear Equations Based Mechanism

Our goal is to design the probability matrix $\mathbf{P} \in [0, 1]^{m \times m}$ whose element is $p_{ij} = \Pr(y = j | x = i)$ such that the utility function $f(p)$ is approximately optimized and we first focus on $f(p) = \text{Error}_{\text{range}}$. Intuitively, in order to obtain optimal utility, a part of probability ratios p_{ik}/p_{jk} would reach to the privacy constraints due to the privacy-utility tradeoff, i.e., inequality constraint function $g_{ij}^{(k)}(p) = 0$ for a subset of $\{i, j, k \in \mathcal{I}\}$ in optimization problem (8). An intuitive idea is to make the p_{kk} as large as possible; thus we let $p_{kk} = e^{\epsilon d_{kj}} \cdot p_{jk}$, i.e., $g_{kj}^{(k)}(p) = 0$, then

$$p_{jk} = e^{-\epsilon d_{kj}} p_{kk}, \quad \forall j, k \in \mathcal{I} \tag{9}$$

After combining equations in (9) and another set of equations $\sum_{k=1}^{m} p_{jk} = 1$ $(\forall j \in \mathcal{I})$, we get the following linear equations

$$\mathbf{E}\mathbf{p} \triangleq \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \cdots & e_{mm} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{22} \\ \vdots \\ p_{mm} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{10}$$

where $e_{jk} \triangleq e^{-\epsilon d_{jk}}$, and $e_{kk} = 1$ because of $d_{kk} = 0$ $(\forall k)$. Since $d_{jk} = d_{kj}$, matrix $\mathbf{E}$ is symmetric. Denote vector $\mathbf{1} = [1, 1, \cdots, 1]^{\mathrm{T}}$, then the linear equations (10) can be rewritten as $\mathbf{E}\mathbf{p} = \mathbf{1}$. When $\mathbf{E}$ is an invertible matrix, and the elements of $\mathbf{p} = \mathbf{E}^{-1}\mathbf{1}$ are non-negative, we can get the values of p_{kk} for all $k \in \mathcal{I}$. Finally, the remaining probabilities p_{jk} $(j \neq k)$ can be calculated by (9).

Note that equations in (9) only guarantee the privacy constraint $p_{ik}/p_{jk} \leqslant e^{\epsilon d_{ij}}$ for $i = k$, not all possible i, j, k. Thus, it is necessary to show that the privacy constraint is satisfied for all $i, j, k \in \mathcal{I}$ as well.

Proposition 1 *If the solutions of linear equations (10) are non-negative, then the probability matrix* $\mathbf{P} = [p_{ij}]_{m \times m}$ *solved by (10) and (9) satisfies local d-privacy with any distance metric.*

Proof: When the solutions $p_{kk} \geqslant 0$ $(\forall k \in \mathcal{I})$ in (10), we have $p_{jk} = e^{-\epsilon d_{jk}} \cdot p_{kk} \geqslant 0$, then $0 \leqslant p_{jk} \leqslant \sum_{k=1}^{m} p_{jk} \leqslant 1$ $(\forall j, k \in \mathcal{I})$. Considering the triangle inequality of the distance metric in (3), we have $d_{jk} - d_{ik} \leqslant d_{ij}$, then the probability ratio is constrained by

$$\frac{p_{ik}}{p_{jk}} = \frac{p_{kk}}{p_{jk}} \cdot \frac{p_{ik}}{p_{kk}} = e^{\epsilon(d_{jk} - d_{ik})} \leqslant e^{\epsilon d_{ij}} \quad (\forall i, j, k) \tag{11}$$

which means the proposed mechanism satisfies local d-privacy with any distance metric. ■

In some cases, the precondition that solutions of (10) are non-negative is not satisfied for very small ϵ, which might be caused by the strong proportional setting in (9), where a smaller ϵ leads to a larger $e^{-\epsilon d_{kj}}$ then a larger p_{kk}, while the summation of some probabilities equals to 1. In the simulation, we observe that for two-dimensional data with Euclidean distance, this precondition is not satisfied for some ϵ less than 1, but for one-dimensional data with absolute value distance, this precondition is still satisfied even for ϵ less than 10^{-5}. Considering that the privacy budget ϵ is usually not very small in practice to obtain reasonable utility, our mechanism is applicable to most practical cases, and how to improve this mechanism to accommodate smaller ϵ will be our future work.

The proportional relationship in (9) seems very similar to the Exponential Mechanism [13], but there is an important difference. Eq. (9) focuses on the probabilities related to the same output, while the proportional relationship in Exponential Mechanism focuses on the probabilities with the same input. The Exponential Mechanism is more general than the proposed one because it can handle any real-valued score functions.

Note that our mechanism can be easily extended to continuous data via discretization. It is applicable to both types of queries including individual queries (such as co-location queries and POI search) and aggregate queries (such as frequency estimation and mean estimation). For POI (points of interest) queries where the user uploads a perturbed location, the server can directly return the list of POIs that are related to this location with good performance because the output of our mechanism is close to the input with high probability, where the perturbation probability is exponentially decreased with the distance according to (9). For mean estimation, the server can utilize the frequency estimation results to estimate the mean of a large number of values with good performance due to the unbiased property of frequency estimation.

B. Optimality for a Specific Range Query

In this part, we show the optimality of the proposed mechanism for a specific range query (i.e., co-location query) via Karush–Kuhn–Tucker (KKT) Conditions [20], which provides the necessary conditions of an optimal solution for the constrained optimization problem. Furthermore, KKT Conditions are also sufficient for an optimal solution when the optimization problem is convex. We omit the details of KKT Conditions in this paper due to the limited pages.

Assume range size $r = 0$ for range queries, which corresponds to the co-location queries. When raw data are distributed uniformly, minimizing $\text{Error}_{\text{range}}$ in (6) is equivalent to minimizing the following objective function

$$f(p) \triangleq -\sum\nolimits_{k=1}^{m} p_{kk} \tag{12}$$

In this case, we show that the solution of the proposed mechanism satisfies KKT Conditions and then is optimal for this optimization problem.

Theorem 1 *If symmetric matrix* $\mathbf{E}$ *in (10) is invertible, and the solutions of (10) are non-negative, i.e., elements of vector* $\mathbf{p} \triangleq \mathbf{E}^{-1}\mathbf{1}$ *are non-negative, then the probability matrix* $p^* \in \mathbb{R}^{m\times m}$ *calculated by (10) and (9) is optimal for the optimization problem in (8) with objective function (12).*

Proof: (Sketch) The Lagrangian function of optimization problem (8) can be represented by

$$\mathcal{L}(p,\mu,\lambda) = f(p) + \sum_{i,j,k} \mu_{ij}^{(k)} g_{ij}^{(k)}(p) + \sum_{k=1}^{m} \lambda_k h_k(p) \tag{13}$$

where the objective function $f(p)$ is defined by (12). Since $\mathbf{p} = \mathbf{E}^{-1}\mathbf{1}$ is non-negative, then $0 \leqslant p_{ik}^* \leqslant 1$ for $i,k \in \mathcal{I}$ (according to Proposition 1). Let multipliers $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_m]^{\mathrm{T}} = \mathbf{p}$ and

$$\mu_{ij}^{(k)} = \begin{cases} \lambda_j e^{-\epsilon d_{ij}}, & k = i \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $\mu_{ij}^{(k)} \geqslant 0$ $(\forall i,j,k)$. Then, we can validate that $p^*, \mu_{ij}^{(k)}$ and λ_k satisfy the KKT Conditions (detailed derivation is shown in Appendix A). Since p^* satisfies constraints $0 \leqslant p_{ij}^* \leqslant 1$ and the optimization problem is linear and thus convex, then p^* is optimal for the considered optimization problem. ■

Note that the objective function in (12) is only a particular case for range query. We will validate the performance of our mechanism in general cases via simulation results.

V. Frequency Estimation and Error Analysis

Since the perturbation matrix of randomize response mechanisms satisfying local d-privacy no longer have the symmetric property like mechanisms under LDP, the frequency estimator in LDP setting is not applicable to our work. Thus, we utilize another unbiased estimator via inversion approach to implement frequency estimation of our mechanism. We analyze the theoretical MSE of this estimator and show the relationship to the estimator in LDP setting.

A. Frequency Estimation

Assume c_k and c_k^* are the reported and true frequency (count) at l_k that is indexed by k

$$c_k = \sum\nolimits_{t=1}^{n} \mathbb{1}_k(y_t), \quad c_k^* = \sum\nolimits_{t=1}^{n} \mathbb{1}_k(x_t)$$

where $x_t, y_t \in \mathcal{I}$ are the mechanism input and output of the user u_t. Indicator function $\mathbb{1}_k(y_t)$ equals to 1 when $y_t = k$, and equals to 0 otherwise. For convenience, denote vectors

$$\mathbf{c} = [c_1, c_2, \cdots, c_m]^{\mathrm{T}}, \quad \mathbf{c}^* = [c_1^*, c_2^*, \cdots, c_m^*]^{\mathrm{T}}$$

For a randomized response based mechanism with probability matrix $\mathbf{P} \in [0,1]^{m\times m}$ whose element is $p_{ij} = \Pr(y = j | x = i)$, assume $\mathbf{P}^{\mathrm{T}}$ is invertible and denote matrix $\mathbf{Q} = (\mathbf{P}^{\mathrm{T}})^{-1}$. The inversion approach [21] considers the following estimator

$$\hat{\mathbf{c}} = (\mathbf{P}^{\mathrm{T}})^{-1}\mathbf{c} = \mathbf{Q}\mathbf{c} \tag{15}$$

We briefly show that $\hat{\mathbf{c}}$ is an unbiased estimator of $\mathbf{c}^*$.

Proposition 2 *For any invertible probability matrix* $\mathbf{P}$*, we have* $\mathbb{E}[\hat{\mathbf{c}}] = \mathbf{c}^*$*, where* $\hat{\mathbf{c}}$ *is defined in (15).*

Proof: The expectation of c_k and $\mathbf{c}$ are

$$\mathbb{E}[c_k] = \sum\nolimits_{i=1}^{m} c_i^* p_{ik} \;\Rightarrow\; \mathbb{E}[\mathbf{c}] = \mathbf{P}^{\mathrm{T}}\mathbf{c}^*$$

Due to the linear property of expectation, we have

$$\mathbb{E}[\hat{\mathbf{c}}] = \mathbb{E}[\mathbf{Q}\mathbf{c}] = \mathbf{Q}\mathbb{E}[\mathbf{c}] = \mathbf{Q}\mathbf{P}^{\mathrm{T}}\mathbf{c}^* = \mathbf{c}^* \tag{16}$$

where $\mathbf{Q}\mathbf{P}^{\mathrm{T}} = \mathbf{I}$ because of $\mathbf{Q} = (\mathbf{P}^{\mathrm{T}})^{-1}$. ■

B. Error Analysis of Frequency Estimation

For convenience, denote the elements of $\mathbf{Q}$ as q_{ij} $(\forall i, j \in \mathcal{I})$. Since $\hat{\mathbf{c}} = \mathbf{Q}\mathbf{c}$, the element of $\hat{\mathbf{c}}$ can be calculated by

$$\hat{c}_k = \sum_{i=1}^{m} q_{ki} c_i = \sum_{i=1}^{m} q_{ki} \sum_{t=1}^{n} \mathbb{1}_i(y_t) = \sum_{t=1}^{n}\sum_{i=1}^{m} q_{ki}\mathbb{1}_i(y_t)$$

Denote $z_k^{(t)}$ as a function of the random variable y_t

$$z_k^{(t)} = \sum\nolimits_{i=1}^{m} q_{ki}\mathbb{1}_i(y_t) \tag{17}$$

then $\hat{c}_k = \sum_{t=1}^{n} z_k^{(t)}$. Considering $y_1, y_2, \cdots, y_n$ are independent because each user independently perturbs her data, then $z_k^{(1)}, z_k^{(2)}, \cdots, z_k^{(n)}$ are independent. The MSE of $\hat{c}_k$ in (7) can be rewritten as

$$\mathrm{MSE}_{\hat{c}_k} = \mathrm{Var}(\hat{c}_k) = \mathrm{Var}\left(\sum_{t=1}^{n} z_k^{(t)}\right) = \sum_{t=1}^{n}\mathrm{Var}(z_k^{(t)}) \tag{18}$$

where $\hat{c}_k$ is an unbiased estimator of c_k^*, and $\mathrm{Var}(z_k^{(t)})$ can be interpreted as the MSE that the user u_t contributes to $\mathrm{MSE}_{\hat{c}_k}$.

According to the definition in (17), the possible values of random variable $z_k^{(t)}$ are $\{q_{k1}, q_{k2}, \cdots, q_{km}\}$ with the probability $\Pr(z_k^{(t)} = q_{ki}) = \Pr(y_t = i | x_t) = p_{x_t i}$. Thus

$$\mathrm{Var}(z_k^{(t)}) = \mathbb{E}[(z_k^{(t)})^2] - \mathbb{E}^2[z_k^{(t)}] = \sum\nolimits_{i=1}^{m} q_{ki}^2 p_{x_t i} - \mathbb{1}_k(x_t)$$

where $\mathbb{E}[z_k^{(t)}] = \sum_{i=1}^{m} q_{ki} p_{x_t i} = \mathbb{1}_k(x_t)$ because of $\mathbf{Q}\mathbf{P}^{\mathrm{T}} = \mathbf{I}$. Since $p_{x_t i} = \sum_{j=1}^{m} \mathbb{1}_j(x_t) p_{ji}$, then the MSE of estimator $\hat{c}_k$ in (18) can be calculated by

$$\begin{aligned}
\mathrm{MSE}_{\hat{c}_k} &= \sum_{t=1}^{n}\mathrm{Var}(z_k^{(t)}) = \sum_{t=1}^{n}\sum_{i=1}^{m} q_{ki}^2 p_{x_t i} - \sum_{t=1}^{n}\mathbb{1}_k(x_t) \\
&= \sum_{j=1}^{m}\left(\sum_{t=1}^{n}\mathbb{1}_j(x_t)\sum_{i=1}^{m} q_{ki}^2 p_{ji}\right) - \sum_{t=1}^{n}\mathbb{1}_k(x_t) \\
&= \sum_{j=1}^{m}\left(c_j^*\sum_{i=1}^{m} q_{ki}^2 p_{ji}\right) - c_k^*
\end{aligned} \tag{19}$$

In this paper, we consider the overall MSE as the utility of frequency estimation, defined by

$$\mathrm{MSE}_{\text{freq}} = \sum_{k=1}^{m}\mathrm{MSE}_{\hat{c}_k} = \sum_{k=1}^{m}\sum_{j=1}^{m}\left(c_j^*\sum_{i=1}^{m} q_{ki}^2 p_{ji}\right) - n \tag{20}$$

where $\sum_{k=1}^{m} c_k^* = n$. We can get the closed-form expressions (elements of matrix $\mathbf{Q}$) for some special cases where the matrix has certain structure, and how to solve it without closed-form expression will be dealt with in future works.

C. Relationship to the Estimator under LDP

The frequency estimator presented in [6] is only applicable to mechanisms with symmetric probability (two different probability values) in the LDP setting, while the estimator in (15) can be used for mechanisms with arbitrary invertible probability matrix $\mathbf{P}$. We show that the estimator in [6] is a reduced case of estimator in (15) under LDP.

When $d_{ij} = 1$ for any $i \neq j$, the notion of local d-privacy reduces to LDP. In this case, the elements of perturbation probability matrix $\mathbf{P} = [p_{ij}]_{m\times m}$ and $\mathbf{Q} = (\mathbf{P}^{\mathrm{T}})^{-1} = [q_{ij}]_{m\times m}$ can be represented as

$$p_{ij} = \begin{cases} a, & i = j \\ b, & \text{otherwise} \end{cases} \qquad q_{ij} = \begin{cases} \frac{1-b}{a-b}, & i = j \\ \frac{-b}{a-b}, & \text{otherwise} \end{cases} \tag{21}$$

where $a \neq b$ and $\sum_{j=1}^{m} p_{ij} = a + (m-1)b = 1$. In (21), the formula of q_{ij} is deducted in Appendix B from the representation of p_{ij}.

According to (15), the unbiased estimator of c_k^* is

$$\hat{c}_k = \sum_{i=1}^{m} q_{ki} c_i = \frac{(1-b)\cdot c_k - b\cdot\sum_{i\neq k} c_i}{a-b} = \frac{c_k - b\cdot n}{a-b}$$

which is identical to the one proposed in [6] for LDP setting, i.e., the estimator in [6] is a special case of the one in (15). Then the MSE of frequency estimation defined in (20) can be calculated by

$$\mathrm{MSE}_{\text{freq}} = \frac{nmb(1-b)}{(a-b)^2} + \frac{n(1-a-b)}{a-b} \tag{22}$$

where a larger domain size m would lead to a larger $\mathrm{MSE}_{\text{freq}}$, which is caused by the privacy constraint of worst-case in LDP. However, for mechanisms satisfying local d-privacy with distance metric, the domain size has little influence on the utility, which can be observed in the simulation.

VI. Evaluation

In this section, we numerically validate the effectiveness of our mechanism and the correctness of our theoretical error analysis by synthetic data and real-world datasets. The following five mechanisms are considered for comparison:

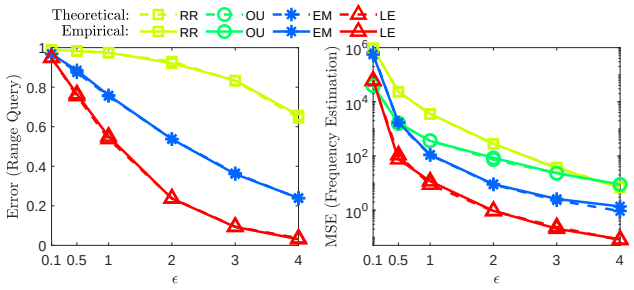
`RR`: generalized Random Response [6], satisfying LDP. The perturbation probabilities are

$$p_{ij} = \Pr(y = j | x = i) = \begin{cases} \frac{e^\epsilon}{e^\epsilon + m + 1}, & i = j \\ \frac{1}{e^\epsilon + m + 1}, & \text{otherwise} \end{cases}$$

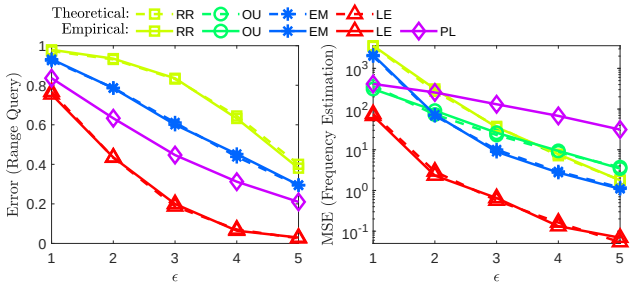
`OU`: Optimized Unary Encoding Mechanism [6], satisfying LDP. The input x is encoded into m bits $B_x = [0, \cdots, 0, 1, 0, \cdots, 0]$ (only the x-th bit is 1), then each bit is perturbed independently with the probabilities

$$\begin{aligned}
\Pr(B_y[i] = 1 | B_x[i] = 1) &= 1/2 \\
\Pr(B_y[i] = 1 | B_x[i] = 0) &= 1/(e^\epsilon + 1)
\end{aligned}$$

Due to the unary encoding, the output B_y might correspond to multiple values, thus is not applicable to answering individual queries such as range queries. We only evaluate its utility for frequency estimation.



(a) Varying ϵ with fixed $m = 100$ (one-dimentional data)



(b) Varying ϵ with fixed $m = 100$ (two-dimentional data)

Fig. 1. Comparisons of theoretical and empirical results

PL: Planar Laplace Mechanism [9], satisfying local d-privacy only with Euclidean distance metric (i.e., geo-indistinguishability). Since it is not based on randomized response type of perturbation (not applicable to the estimator in (15)), we use the raw frequency (without estimator) and only show its empirical performances for two-dimensional data.

EM: Exponential Mechanism [13], which designs the perturbation probability by

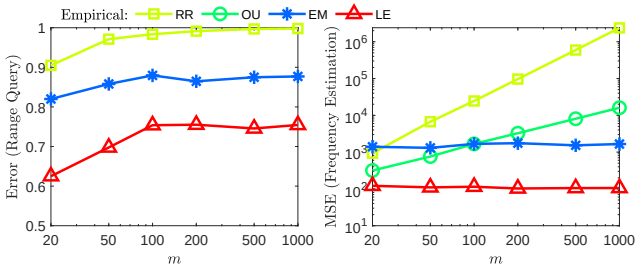
$$p_{ij} = \Pr(y = j | x = i) \propto \exp(\epsilon/2 \cdot \mathcal{Q}(x, y))$$

where $\mathcal{Q}(x, y)$ is a score function to quantify the desirability of outputs. We utilize $\mathcal{Q}(x, y) = -d(x, y)$ in order to satisfy local d-privacy.

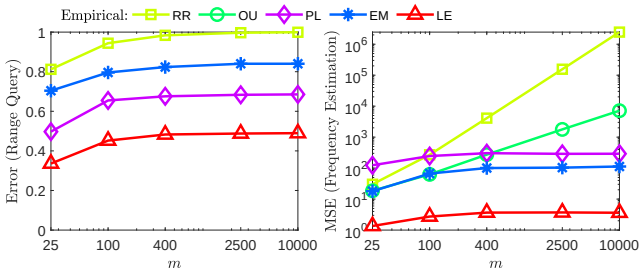
LE: Linear Equations Mechanism (the proposed one), satisfying local d-privacy with any distance metric. It might not obtain the optimal solution of the optimization problem (12) in general case, but it can reach the optimum for some specific objective functions (refer to Section IV-B).

Considering that the above mechanisms satisfy different privacy notions, the distance metric is normalized by $d_{ij} \leftarrow d_{ij}/d_{\min}$, where $d_{\min}$ is the minimal distance of $i, j \in \mathcal{I}$, and in the case of location it represents the size of the grid that divides the area. Then, local d-privacy is a relaxed version of LDP because of $d_{ij} \geqslant 1$ $(\forall i, j \in \mathcal{I})$.

The evaluation metrics are the Error of range query defined in (6) and the MSE of frequency estimation defined in (20). Since MSE_{freq} in (20) is proportional to the number of users, we scale the metric via $\text{MSE}_{\text{freq}} \leftarrow \text{MSE}_{\text{freq}}/n$ in evaluation, where n is the number of users. Note that all of the considered mechanisms are independent of evaluation metrics, thus we show the two utilities separately in order to clearly observe the performance of these mechanisms.



(a) Varying m with fixed $\epsilon = 0.5$ (one-dimentional data)



(b) Varying m with fixed $\epsilon = 2$ (two-dimensional data)

Fig. 2. Influence of domain size m on utilities

A. Synthetic Data

We consider two formats of synthetic data: one-dimensional and two-dimensional. For one-dimensional data, the domain is $\mathcal{D} = \{1, 2, \cdots, m\}$ (identical to the index set $\mathcal{I}$), and the distance metric is $d_{ij} = |i - j|$. For two-dimensional data, assume $\sqrt{m}$ is an integer for simplicity. Then the data domain is $\mathcal{D} = \{(i_1, i_2)\}_{i_1, i_2 = 1}^{\sqrt{m}}$ (the domain size is still m), and the distance metric is $d_{ij} = \sqrt{(i_1 - j_1)^2 + (i_2 - j_2)^2}$ (Euclidean distance). In the two cases, the number of users is $n = 10 \cdot m$ and the empirical results are averaged with five repeats.

Validation of theoretical analysis. Considering that the data distribution is often uneven and sparse, we randomly generate the raw data with random distribution on only 20% of the domain. The theoretical and empirical results for both one-dimensional and two-dimensional data are compared in Fig. 1. The *empirical* utilities are computed as $\text{Error}_{\text{range}} = 1 - \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{x_t}(y_t)$ (with range size $r = 0$) and $\text{MSE}_{\text{freq}} = \frac{1}{n}\sum_{k=1}^{m}(\hat{c}_k - c_k^*)^2$. For two-dimensional data, the ϵ starts from 1 because the precondition of LE (solutions of linear equations (10) are non-negative) is not satisfied for some ϵ less than 1 in this Euclidean distance case (explained in Section IV-A). The difference between theoretical and empirical results are almost negligible in Fig. 1, which validates the correctness of our theoretical error analysis. In the following simulation, we only show the empirical results for simplicity.

Influence of domain size. The simulation results with different domain size m are shown in Fig. 2, where the MSE_{freq} of RR and OU (satisfying LDP) is proportional to m. But for PL, EM and LE (satisfying local d-privacy), the domain size m has relatively less influence on the performance. Generally speaking, the privacy definition with distance metric relaxes LDP notion and improves utility, especially for large-scale data domain. Under the considered ϵ and m in Fig. 1 and Fig. 2, the

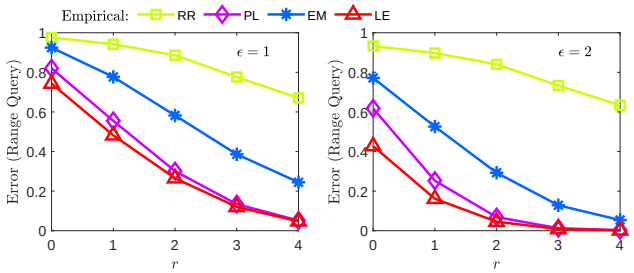


Fig. 3. Varying r of range query when $m = 100$ (two-dimensional data)

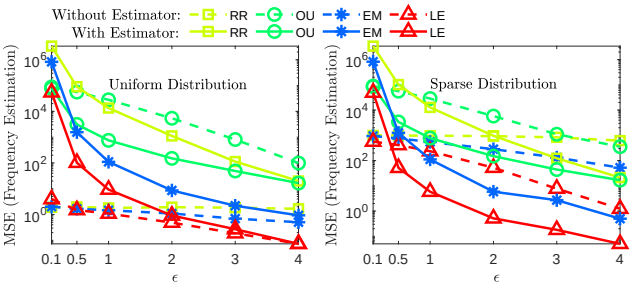


Fig. 4. Comparisons of MSE_{freq} under two types of distributions when $m = 200$ (one-dimensional data). Under the uniform distribution, user's raw data are uniformly distributed in the domain. While under the sparse distribution, user's raw data are distributed only on 1% of the domain.

proposed LE has the best or near-best performance for both range query and frequency estimation. Since the proportional relationship in EM focuses on probabilities with the same input, it does not directly guarantee privacy constraints, thus needs a factor $\frac{1}{2}$ for privacy guarantees, which deteriorates the utility.

Influence of range size for range query. In previous cases, we let $r = 0$ in range query for simplicity, which corresponds to co-location query. Fig. 3 shows the $\text{Error}_{\text{range}}$ when r is greater than 0. Since distance d_{ij} is normalized by d_{ij}/d_{min}, the range size r is normalized as well. For the mechanisms satisfying local d-privacy with distance metric (e.g., PL, EM and LE), a larger r for range query would reduce the $\text{Error}_{\text{range}}$ significantly because the raw data are perturbed to nearby values (with small distance) with higher probabilities.

Benefit of estimator for frequency estimation. In the previous simulation, mechanisms satisfying local d-privacy adopted the unbiased estimator $\hat{\mathbf{c}} = \mathbf{Q}\mathbf{c}$ in (15) for frequency estimation, and mechanisms with LDP guarantees (RR and OU) adopted the estimator presented in [6], which is a special case of the estimator in (15). However, the unbiasedness does not always lead to the minimum MSE. The MSEs of $\mathbf{c}$ (without estimator) and $\hat{\mathbf{c}}$ (with estimator) under two distributions are compared in Fig. 4. For OU with unary encoding, the single output vector might correspond to multiple values; thus the unbiased estimator can efficiently reduce the MSE. While for other mechanisms with direct encoding, an output only corresponds to one value. Without the estimator, the performance significantly depends on the distribution of data. But for mechanisms with estimator, the data distribution has relatively less influence on the performance. Generally speaking, the frequency estimation method with estimator is superior to that without estimator for two reasons: (i) The data distribution has less impact on the MSE, thus is more stable and reliable. (ii)

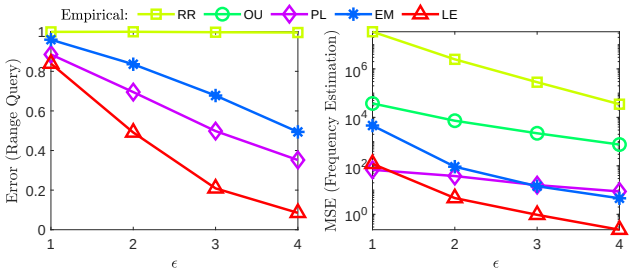


Fig. 5. Empirical results of real-world data ($m = 10,000$)

It has better performance for sparse distribution, which is very common in practice.

B. Real-World Data

We compare the performances of these mechanisms with the real-world dataset Gowalla [22], which is available in [23]. There are 196,585 users who share their locations (latitude and longitude) by checking-in, and 6,442,890 check-ins are recorded. For simplicity, we consider the locations that are first checked-in by users within an area where the latitude is from 50 to 55 and the longitude is from 0 to -5 (this area covers most of the United Kingdom). We uniformly divide this area into 100×100 districts by latitude and longitude. Each user's check-in location corresponds to one of these districts, where majority of the districts have few users and approximately 25% users are distributed in the top-5 popular districts. The utility is shown in Fig. 5, where the privacy and utility definitions are the same as the settings in the synthetic two-dimensional data. The performances are very similar to the previous results, where LE outperforms all other mechanisms in both range queries and frequency estimation queries and the utility of mechanisms satisfying LDP deteriorate extremely due to the large domain size.

VII. Conclusion and Future Work

In this paper, a new mechanism is proposed to answer range queries and provide frequency estimation with good performance at the same time while satisfying local d-privacy with any distance metric. It is proved to have optimal utility for co-location query (a specific type of range query). It can be applied to frequency estimation with an unbiased estimator. We validate the effectiveness of our mechanism and correctness of the theoretical MSE analysis via synthetic data and real-world dataset. The simulation results show the advantage of our mechanism and the mechanisms satisfying local d-privacy, compared with the ones satisfying LDP.

For future work, we will extend our work to handle more types of queries and general utility functions, such as heavy hitter estimation. We will improve our mechanism to obtain better properties for frequency estimation.

References

[1] C. Dwork, "Differential privacy," in *ICALP*, 2006, pp. 1–12.
[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference (TCC)*, 2006, pp. 265–284.

[3] R. Chen, H. Li, A. Qin, S. P. Kasiviswanathan, and H. Jin, "Private spatial data aggregation in the local setting," in *IEEE International Conference on Data Engineering (ICDE)*, 2016, pp. 289–300.
[4] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *ACM Conference on Computer and Communications Security (CCS)*, 2014, pp. 1054–1067.
[5] "Learning with privacy at scale," https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html, 2017.
[6] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proceedings of the 26th USENIX Security Symposium*, 2017, pp. 729–745.
[7] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *ACM Symposium on Theory of Computing (STOC)*, 2015, pp. 127–135.
[8] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Privacy Enhancing Technologies*, 2013, pp. 82–102.
[9] M. Andrés, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *20th ACM CCS'*, 2013, pp. 901–914.
[10] H. Pham, C. Shahabi, and Y. Liu, "Ebm: an entropy-based model to infer social strength from spatiotemporal data," in *ACM SIGMOD International Conference on Management of Data*, 2013.
[11] C. Shahabi, L. Fan, L. Nocera, L. Xiong, and M. Li, "Privacy-preserving inference of social relationships from location data: a vision paper," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015.
[12] 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.
[13] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2007, pp. 94–103.
[14] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Geo-indistinguishability: A principled approach to location privacy," in *International Conference on Distributed Computing and Internet Technology*, vol. 8956, 2015, pp. 49–72.
[15] 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 *ACM CCS'*, 2016, pp. 192–203.
[16] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, "Cap: A context-aware privacy protection system for location-based services," in *2009 29th IEEE International Conference on Distributed Computing Systems*. IEEE, 2009, pp. 49–57.
[17] B. Jiang, M. Li, and R. Tandon, "Context-aware data aggregation with localized information privacy," in *IEEE Conference on Communications and Network Security (CNS)*, 2018, pp. 1–9.
[18] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *ACM SIGSAC CCS'*, 2015, pp. 1298–1309.
[19] W. Zhang, M. Li, R. Tandon, and H. Li, "Online location trace privacy: An information theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 235–250, 2019.
[20] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
[21] Z. Huang and W. Du, "Optrr: Optimizing randomized response schemes for privacy-preserving data mining," in *IEEE International Conference on Data Engineering (ICDE)*, 2008, pp. 705–714.
[22] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1082–1090.
[23] "Gowalla dataset," https://snap.stanford.edu/data/loc-gowalla.html.
[24] N. J. Higham, *Accuracy and stability of numerical algorithms*. SIAM, 2002, vol. 80.

APPENDIX A
VERIFICATION OF KKT CONDITIONS IN THEOREM 1

Proof: Considering $\mathbf{E}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{E}\mathbf{p} = \mathbf{1}$, we have

$$\lambda_i + \sum\nolimits_{j\neq i} \mu_{ij}^{(i)} = \lambda_i + \sum\nolimits_{j\neq i} \lambda_j e^{-\epsilon d_{ij}} = 1 \ (\forall i) \qquad (23)$$

According to the definition, only the following partial derivatives are non-zero

$$\partial f/\partial p_{kk} = -1, \quad \partial h_k/\partial p_{ki} = 1,$$
$$\partial g_{ij}^{(k)}/\partial p_{ik} = 1, \quad \partial g_{ji}^{(k)}/\partial p_{ik} = -e^{\epsilon d_{ij}} \quad (i \neq j)$$

For $\mathcal{L}(p, \mu, \lambda)$ in (13), the partial derivative $\partial\mathcal{L}/\partial p_{tt}$ at p^* is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p_{tt}} &= -1 + \sum_{j=1}^{m} \mu_{tj}^{(t)} \frac{\partial g_{tj}^{(t)}}{\partial p_{tt}} + \sum_{i=1}^{m} \mu_{it}^{(t)} \frac{\partial g_{it}^{(t)}}{\partial p_{tt}} + \lambda_t \frac{\partial h_t}{\partial p_{tt}} \\ &= -1 + \sum\nolimits_{j\neq t} \mu_{tj}^{(t)} + 0 + \lambda_t = \lambda_t + \sum\nolimits_{j\neq t} \mu_{tj}^{(t)} - 1 \\ &= 1 - 1 = 0 \qquad \text{(according to (23))}\end{aligned}$$

And the partial derivative $\partial\mathcal{L}/\partial p_{st}$ $(s \neq t)$ at p^* is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p_{st}} &= 0 + \sum_{j=1}^{m} \mu_{sj}^{(t)} \frac{\partial g_{sj}^{(t)}}{\partial p_{st}} + \sum_{i=1}^{m} \mu_{is}^{(t)} \frac{\partial g_{is}^{(t)}}{\partial p_{st}} + \lambda_s \frac{\partial h_s}{\partial p_{st}} \\ &= 0 + 0 + \mu_{ts}^{(t)} \frac{\partial g_{ts}^{(t)}}{\partial p_{st}} + \lambda_s = \lambda_s - \mu_{ts}^{(t)} \cdot e^{\epsilon d_{st}} \\ &= \lambda_s - \lambda_s = 0 \qquad \text{(by definition } \mu_{ij}^{(i)} = \lambda_j e^{-\epsilon d_{ij}})\end{aligned}$$

Thus we have $\nabla_p \mathcal{L}(p^*, \mu, \lambda) = \mathbf{0}$. For any $i, j, k \in \mathcal{I}$, the remaining KKT Conditions (refer to [20]) are satisfied as well: $g_{ij}^{(k)}(p^*) \leqslant 0$, $h_k(p^*) = 0$ and $\mu_{ij}^{(k)} \geqslant 0$ are guaranteed by Proposition 1 and definitions; Equation $\mu_{ij}^{(k)} \cdot g_{ij}^{(k)}(p^*) = 0$ is satisfied because of $g_{ij}^{(i)}(p^*) = 0$ (from (9)) and $\mu_{ij}^{(k)} = 0$ when $k \neq i$ or $i = j$. ■

APPENDIX B
THE DEDUCTION OF q_{ij} IN (21)

We first show a useful lemma and then implement the deduction.

Lemma 1 *(Woodbury Matrix Identity [24]) Given matrices $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times k}, \mathbf{C} \in \mathbb{R}^{k\times k}, \mathbf{D} \in \mathbb{R}^{k\times n}$, if both $\mathbf{A}$ and $\mathbf{C}$ are invertible, then*

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1}$$

If $b = 0$, the formula of q_{ij} in (21) is obvious. If $b \neq 0$, let

$$\mathbf{A} = (a-b)\mathbf{I}, \quad \mathbf{C} = b, \quad \mathbf{B} = \mathbf{D}^{\mathrm{T}} = [1, 1, \cdots, 1]_{m\times 1}$$

where $\mathbf{I}$ is identity matrix with size $m \times m$. Considering that the element p_{ij} of $\mathbf{P}$ has the format in (21), we have

$$\mathbf{P}^{\mathrm{T}} = \mathbf{P} = (a-b)\mathbf{I} + b \cdot \mathbf{BB}^{\mathrm{T}} = \mathbf{A} + \mathbf{BCD}$$

where $a \neq b$ and $a + (m-1)b = 1$, then

$$\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1} = \frac{m}{a-b} + \frac{1}{b} = \frac{a + (m-1)b}{(a-b)b} = \frac{1}{(a-b)b}$$

According to Woodbury Matrix Identity, we have

$$\begin{aligned}\mathbf{Q} = (\mathbf{P}^{\mathrm{T}})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1} \\ &= \mathbf{A}^{-1} - (a-b)b \cdot \mathbf{A}^{-1}\mathbf{BDA}^{-1} = \frac{1}{a-b}\left(\mathbf{I} - b \cdot \mathbf{BD}\right)\end{aligned}$$

Thus we get the formula of q_{ij} in (21).