K-means clustering with local d_χ-privacy for privacy-preserving data analysis

Mengmeng Yang*, *Member, IEEE,* Ivan Tjuawinata*, *Member, IEEE,* Kwok-Yan Lam, *Senior Member, IEEE*

***Abstract*—Privacy-preserving data analysis is an emerging area that addresses the dilemma of performing data analysis on user data while protecting users' privacy. In this paper, we consider the problem of constructing privacy-preserving K-means clustering protocol for data analysis that provides local privacy to users' data. To enable a desirable degree of local privacy guarantee while maintaining high accuracy of the clustering, we adopt a generalized differential privacy definition, d_χ-privacy, which quantifies the distinguishability level based on the distance between data records defined by the distance function d_χ. In our work, we consider the space of data points as a metric space imbued with Euclidean distance and propose a bounded perturbation mechanism (BPM) with bounded sampling space of the perturbed data points, which is formally shown to achieve d_χ-privacy. BPM perturbs the data as a whole instead of treating each dimension independently, which is desirable since the privacy budget is no longer required to be split among different dimensions. Bounded output space also means that we will not get into the case where the report or the statistical result is so far out of the data domain that it is hard to interpret. Furthermore, it can also help in limiting the amount of bandwidth needed to send such report to the server. The design of BPM is based on a probability density function which decreases exponentially as the Euclidean distance with respect to the true value grows. It is also designed with the aim of ensuring that BPM produces perturbed data that provides the claimed privacy guarantee while ensuring high utility response. To guarantee the efficiency of the perturbation method, we propose an efficient algorithm to sample from the proposed distribution and apply BPM to the design of d_χ-private K-means clustering algorithms. Lastly, we analyse the privacy and utility guarantee provided by the proposed method and provide its experimental results.**

***Index Terms*—Differential privacy, K-means clustering, d_χ-privacy.**

I. Introduction

With the emergence of internet application services, huge amount of data are generated and collected everyday. In order to unlock the value of such data, data analytics have emerged as a strategic research area to enhance competitiveness in cyberspace. However, the task of performing effective data analysis has become increasingly challenging due to the enactment of data privacy regulations in many jurisdictions. K-means clustering, as one of the most popular unsupervised learning algorithms, has been widely used in many data-driven applications, such as document clustering [1], [2], spam filtering [3], anomaly detection [4], and customer segmentation [5]. The target of the algorithm is to partition the data points into different groups according to feature similarity such that the data points in the same group are closer to each other and are further to data points from other groups. In this paper, we consider privacy-preserving K-means clustering scheme that aims to protect users' data *locally* while enabling high quality data analysis.

Local differential privacy (LDP) has been recently deployed by many big companies [6], [7] to collect and analyse their users' data due to its rigorous privacy guarantee. In LDP, each user perturbs their data locally and only sends the perturbed value to the server. The privacy guarantee is defined by giving some upper bound on the statistical difference between the distribution of the perturbed responses based on any two possible data records in the domain, also known as the distinguishability level. In LDP, the distinguishability level is set to be the same between any pair of data record. Intuitively, this provides the same level of privacy guarantee for user's sensitive data with respect to any other data. One property that LDP has is the relatively larger accumulated variance from different reports. Therefore, the majority of the state-of-the-art LDP mechanisms only works well for simple statistical computations, such as mean value estimation [8], and uni-dimensional data with large number of population. To enable a desirable degree of local privacy guarantee while maintaining high accuracy of the clustering, we adopt a generalized differential privacy definition, d_χ-privacy.

The concept of d_χ privacy is defined for data in a metric space equipped with a distance function d_χ, such as Manhattan distance, Euclidean distance, and Chebyshev distance. Given such data space with an embedded distance function, d_χ-privacy is defined to have the requirement for the distinguishability level of two data to be based on the distance between the two data records. In this work, the data space is defined to be the d-dimensional real-valued space $\mathbb{R}^d$ imbued with the L^2-norm, which is better known as Euclidean norm. To the best of our knowledge, this is the first work that considers the problem of K-means clustering with d_χ-privacy.

Specifically, we propose a bounded perturbation mechanism, which guarantees d_χ-privacy for each user. In contrast to general mechanisms, BPM perturbs the data as a whole instead of treating each dimension independently, which is desirable since the privacy budget is no longer required to be split among different dimensions. Furthermore, BPM outputs a perturbed response that is within a bounded domain, which guarantees

M. Yang, I. Tjuawinata, and K-Y. Lam are with the Strategic Centre for Research in Privacy-Preserving Technologies & Systems, Nanyang Technological University, Singapore.

E-mail: {melody.yang, ivan.tjuawinata, kwokyan.lam}@ntu.edu.sg

*Both authors contributed equally to this research.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes proofs of some of the claims that are omitted from the main manuscript and some additional experimental results. Contact melody.yang@ntu.edu.sg or ivan.tjuawinata@ntu.edu.sg for further questions about this work.

that we can have a reasonable interpretation of the mechanism output which may also be necessary for further analysis and decision making while also limiting the amount of bandwidth needed to send such report to the server.

The design of such bounded mechanism is nontrivial. Firstly, the space where the perturbed value is sampled needs to be designed carefully. In order to preserve privacy guarantee, such space needs to be independent of the actual value of the private data. Furthermore, we want the distribution of such value to be positively correlated with its distance from the actual value. To meet the above requirements, we design such space to contain the domain of the actual private data as its subspace. Secondly, to satisfy the privacy definition, the ratio of two distribution of the perturbed responses needs to be bounded. To provide such a upper bound, we define the distribution of our mechanism based on two distributions. More specifically, we design the density function to decrease as the distance to the actual value increases from 0 to a predetermined and public constant L and the density function becomes constant as the distance increases from L. Such constant is determined based on other publicly known information such as the privacy budget and the data domain and is independent of the actual values of the private data.

Overall, our main contributions are summarized as follows.

- We propose a Bounded Perturbation Mechanism (BPM) and formally prove that the proposed BPM satisfies d_χ-privacy in terms of Euclidean distance. The report obtained through BPM is shown to have a small asymptotic expected distance to the real data, which is shown to be $O(d)$, where d is the data dimension. Furthermore, as our theoretical analysis suggests, although the mechanism is not unbiased, asymptotically, it provides a smaller variance with respect to the data dimension.
- We applied BPM to design a d_χ-private K-means clustering algorithm to protect the user's privacy locally. We further propose an efficient sampling algorithm that realizes the proposed BPM with the desired probability density function under the assumption that an efficient one-dimensional uniform sampling algorithm exists.
- We provide theoretical analysis on the privacy as well as the utility of our proposed method. This analysis fills the gap in theoretical studies of locally private K-means clustering algorithms with d_χ-privacy protection. Furthermore, we conducted extensive experiments to show the performance of our proposed method.

The rest of this paper is organized as follows. Section II introduces the preliminaries. The proposed solution is proposed in Section III. Section IV and Section V show the experimental result and related work respectively while the paper is concluded in Section VI.

II. Preliminaries

A. Problem Definition

Let U be a set of users where each user $u_i \in U$ holds a multi-dimensional data $\mathbf{v}_i \in \mathbb{R}^d$ where d is the number of attributes of the data. Given such data, the server is interested to generate a partition of the users to K clusters for some positive integer K through the use of a K-means clustering algorithm. In order to provide some privacy protection to the user's data, each user performs some randomization and perturbation to its data before reporting it to the server.

TABLE I
Notations

Notation	Description
X	data set
u_i	the i-th user
$\mathbf{v}_i$	data record of u_i
$\mathbf{x}_i$	perturbed data record of u_i
$v_{i,j}$	the value for attribute j of $\mathbf{v}_i$
d	data dimension
$d(\cdot,\cdot)$	distance function
k	privacy parameter
K	the number of clusters
$\mathbb{C}_t$	t-th cluster
$\mathbf{c}_t$	centroid of cluster $\mathbb{C}_t$
ϵ	privacy budget
$\mathcal{M}$	randomized algorithm
r, R	radius
L	constant

In this study, our aim is to develop a perturbation mechanism that may (a) prevent other parties from learning significant information about the user's private data from his response while also (b) producing "accurate" clustering which is comparable to that generated in the clear. In our work, we are assuming the existence of an untrusted server that aims to calculate some function based on the private data owned by users. In our scenario, in order to protect the privacy of each user's data, each user works independently of each other to perturb his private data before reporting it directly to the server. Due to the independence of different users, no information is ever sent from one user to another. Furthermore, note that such mechanism only contains one round of communication where each user sends a perturbed report to the server. Since no private information is ever sent to any user, an adversary corrupting a set of users can never learn any additional information on private data held by other users. Lastly, we assume that the adversary is semi-honest, i.e., it learns all the information held by corrupted parties, tries to learn more information about private data, but it must follow the protocol honestly. Hence, it is impossible for the adversary to affect the computation of any mechanism in our protocol. This implies that an adversary can gain nothing from corrupting any user except for completely learning the private value of such corrupted user. Because of this, in our work, we always assume that the semi-honest adversary may only control the server, which allows it to learn the perturbed reports of all the users and it tries to learn more information regarding the users' private data from those reports. For reference, Table I contains some common notations used throughout the work.

B. K-means Clustering

K-means clustering algorithm is one of the most popular unsupervised machine learning algorithms to partition data set to clusters. The target of K-means algorithm is to classify data points into several groups maximizing the similarity between

any data in the same group while minimizing the similarities of data across distinct groups. In this work, we consider a basic K-means clustering algorithm, which is known as the Lloyd's algorithm [9].

Formally, we are given a set of observations $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ where each observation is a d-dimensional feature vector we denote by $\mathbf{v}_i = (v_{i,1}, \cdots, v_{i,d})$. K-means clustering algorithm aims to partition the n data to K clusters $\mathbb{C}_1, \cdots, \mathbb{C}_K$ such that $\bigcup_{i=1}^{K} \mathbb{C}_i = \{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and it minimizes the sample variance for each cluster. Let $C = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ be the corresponding K centroids, which is defined as the average of the points in the corresponding clusters, the problem can then be formulated as the following optimization problem

$$\arg \min_{\mathbb{C}_1, \cdots, \mathbb{C}_K \text{ partition of } \{\mathbf{v}_1, \cdots, \mathbf{v}_N\}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in \mathbb{C}_j} ||\mathbf{x} - \mathbf{c}_j||^2. \quad (1)$$

The specific steps are as follows.

- We initialize a set of K random centroids.
- For each data point, we calculate the distance between the data point and all data centroids. We include such data point to the cluster corresponding to the nearest centroid to obtain K clusters.
- For each cluster, we calculate the average of the data points categorized to such cluster and set the average to be the updated centroid of the cluster.
- We reiterate the second and third steps discussed above while the stop condition is not met.

C. d_χ-privacy

In this section, we consider a generalization of the concept of differential privacy over some metric space χ. This generalization, denoted as d_χ-privacy, was first proposed by Chatzikokolakis *et al.* [10] in 2013. Formally, d_χ is defined as follows.

Definition 1. *Let χ be a metric space containing all possible values of the user's data and $\mathcal{R}$ be a report space containing all possible reports the user may send to the server. Let $\mathcal{M} : \chi \to \mathcal{R}$ be a randomized mechanism a user may use to generate a report based on his data. Then $\mathcal{M}(\mathbf{x})$ is a random variable following the distribution defined by the probability density function we denote by $f_{\mathbf{x}} : \mathcal{R} \to \mathbb{R}_{\geq 0}$. $\mathcal{M}$ is said to be d_χ-private if and only if for any $\mathbf{x}, \mathbf{x}' \in \chi$ and $\mathbf{y} \in \mathcal{R}$,*

$$f_{\mathbf{x}}(\mathbf{y}) \leq e^{d_\chi(\mathbf{x}, \mathbf{x}')} f_{\mathbf{x}'}(\mathbf{y}).$$

Intuitively, the statistical distance between two distributions outputted by a d_χ-private mechanism with respect to two distinct data decreases as the distance between the two data decreases with respect to the distance function d_χ. In order to incorporate a measure of privacy loss, we may embed the privacy budget to the distance function. More specifically, instead of having d_χ to be the original distance function, we have $d_\chi(\mathbf{x}, \mathbf{x}') = \epsilon d(\mathbf{x}, \mathbf{x}')$ where $d(\cdot, \cdot)$ is the original distance function. Therefore, for a fixed $\epsilon > 0$, a mechanism $\mathcal{M}$ is said to have ϵd privacy if for any $\mathbf{x}, \mathbf{x}' \in \chi$ and $\mathbf{y} \in \mathcal{R}$,

$$f_{\mathbf{x}}(\mathbf{y}) \leq e^{\epsilon d(\mathbf{x}, \mathbf{x}')} f_{\mathbf{x}'}(\mathbf{y}).$$

Note that the definition of d_χ privacy is a generalization of traditional differential privacy notions, both the central and local variants. More specifically, a central differential privacy notion limits the definition of d_χ-privacy to only consider pair of data-sets X and X' such that $d(X, X') = 1$. On the other hand, a local differential privacy notion provides the bound between the two probability density function to be a constant with respect to the distance between the two corresponding data while d_χ-privacy notion generalizes this bound to be a linear function with respect to their distance. This definition also allows privacy notion to be defined in various other metrics which can be more suitable in various applications. Manhattan distance may be more suitable in the scenario where the data domain involves date and time [10] while Chebyshev distance may be more suitable for the study of privacy aspect of readings for smart meters [11]. On the other hand, in textual data, the hyperbolic distance may be more suitable to encode hierarchical and semantic information [12] while Earth Mover's distance is more suitable to measure the distance between different words [13]. Although an instance of d_χ-privacy has been used in [13] under the name of geo-indistinguishability, they only consider data with 2-dimensions. In this work, we consider Euclidean distance to measure the distance between two d-dimensional data records belonging to $\mathbb{R}^d$. Throughout our work, we use the notation $\|\mathbf{x}\|_2$ to represent the Euclidean norm of $\mathbf{x}$ and $d_E(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$ represent the Euclidean distance between the two data records. We refer ϵd_E-privacy to the d_χ-privacy with respect to Euclidean distance.

III. K-means Clustering with ϵ d_E-privacy

In this section, we propose a perturbation method for K-means clustering algorithm providing local ϵd_E-privacy.

A. General Construction

In this section, we provide a generic framework for our proposed privacy-preserving K-means clustering algorithm given a perturbation mechanism $\mathcal{M}$, which can be found in Algorithm 1. Specifically, we let each user perturb his own real data record using a perturbation method $\mathcal{M}$ which is then reported to the server. Having these perturbed records, the server performs a K-means clustering algorithm with such records as the inputs to estimate the clusters.

Compared to an interactive model, a non-interactive model does not need multiple access to the private data, which saves the privacy budget. However, it provides more challenges to the perturbation method since it needs to both provide sufficient privacy guarantee to the user's real data record and enable an effective K-means clustering algorithm. In this work, we model the data domain as a multi-dimensional vector and the perturbation is modelled following a random variable with the original user's data as its centre. For simplicity, we define the data domain as $\mathcal{D} = [0, 1]^d$. In other words, any data is a d-dimensional real-valued vector with each entry being a real number in $[0, 1]$. We note that a simple normalization method can be done to transform any d-dimensional real-valued data to be an element in $\mathcal{D}$.

Algorithm 1 K-means clustering with ϵd_E-privacy

Require: User's data record $\mathbf{v}$ of dimension d
Ensure: K clusters $\mathbb{C}_1, \cdots, \mathbb{C}_K$
1: [*User side:*]
2: **for** each user $u_i \in U$ **do**
3: Execute $\mathbf{x}_i \leftarrow \mathcal{M}(\mathbf{v}_i)$;
4: Report $\mathbf{x}_i$ to the server;
5: **end for**
6: [*Server side:*]
7: Perform K-means algorithm after getting reports $\mathbf{x}_1, \cdots, \mathbf{x}_n$ from all users to obtain the clusters $\mathbb{C}_1, \cdots, \mathbb{C}_K$;
8: **return** The clusters $\mathbb{C}_1, \cdots, \mathbb{C}_K$;

We note that such general framework can be used to obtain a privacy-preserving K-means clustering algorithm given that the mechanism $\mathcal{M}$ also provides the required privacy guarantee. This claim can be easily verified and has been formally shown in Theorem 9. Next we consider a concrete privacy-preserving K-means clustering algorithm based on the framework described in Algorithm 1 by considering a specific ϵd_E-private mechanism $\mathcal{M}$. Note that general perturbation mechanisms such as Gaussian and Laplacian mechanisms have the whole real space as their co-domain. This implies that, although it comes with a very small probability, it is possible for a perturbed report to take any value in the whole real space. This property may not be desirable in all situations. In particular, recall that each of the coordinates of the data in the data domain $\mathcal{D}$ can only take a real number value between 0 and 1. So assuming that we send each data with precision of p bits, such data can be represented as p-bit string. However, by using Gaussian or Laplacian mechanism, the perturbed report can be arbitrarily large, making the amount of data required to be sent to also be arbitrarily large. Furthermore, such perturbed data may be harder to interpret due to it being very far from the original domain. In order to resolve such issues, we consider the use of a perturbation mechanism with bounded sampling space. Specifically, we propose a bounded perturbation mechanism which we denote by BPM, which uses a bounded domain $\mathcal{R} \subseteq \mathbb{R}^d$ as the sample space such that $[0,1]^d \subseteq \mathcal{R}$. In the remainder of this section, we discuss BPM in more detail.

B. Bounded perturbation mechanism

In this section, we propose a perturbation mechanism $\mathcal{M}$ that satisfies the ϵd_E-privacy for any given privacy budget $\epsilon > 0$, that is, for any $\mathbf{v}, \mathbf{v}' \in \mathcal{D}$ and $\mathbf{y} \in \mathcal{R}$,

$$f_{\mathbf{v}}(\mathbf{y}) \leq e^{\epsilon d_E(\mathbf{v},\mathbf{v}')} f_{\mathbf{v}'}(\mathbf{y}).$$

Let L be a positive real number that is publicly known and fixed for all users. We assume the choice of L is done by the server and is independent of the users' private data. Since it is independent of the value of the actual private data and generated by the server itself, the value of L does not leak any information about any private data to the server. The choice of the value L will be discussed in more detail in Remark 2.

Given a data record $\mathbf{v} = (v_1, \cdots, v_d)$, our intention is to sample a noisy vector $\mathbf{x}$ following a specific distribution $f_{\mathbf{v}}^L$, which ensures the noisy record $\mathbf{x}$ is statistically indistinguishable from the original true record $\mathbf{v}$. As discussed above, instead of letting the whole $\mathbb{R}^d$ to be the space of all possible reports, we sample the report from a fixed bounded domain, which is a function of the constant L.

The formal definition is given as follows.

Definition 2 (Bounded Perturbation Mechanism (BPM))**.** *Let k be the privacy parameter, $\mathbf{v} = (v_1, \cdots, v_d) \in \mathcal{D}$ be a d dimensional data record, and L be a positive real number. Let the report space $\mathcal{R}_L \triangleq [-L, 1+L]^d \subseteq \mathbb{R}^d$. We define the bounded perturbation mechanism $\mathcal{M}$ as follows. The random variable $\mathcal{M}(\mathbf{v})$ is defined to be a random variable X following the probability distance function $f_{\mathbf{v}}^{(L)} : \mathcal{R}_L \to \mathbb{R}_{\geq 0}$ such that for any $\mathbf{x} \in \mathcal{R}_L$,*

$$f_{\mathbf{v}}^{(L)}(\mathbf{x}) = \lambda_{\mathbf{v},L} e^{-k(\min\{\|\mathbf{x}-\mathbf{v}\|_2, L\})}, \tag{2}$$

where $\lambda_{\mathbf{v},L}$ is a constant such that $\int\cdots\int_{\mathcal{R}_L} f_{\mathbf{v}}^{(L)}(\mathbf{x}) d\mathbf{x} = 1$.

To get the value of $\lambda_{\mathbf{v},L}$, we define some notations first. Let d be a positive integer, $\mathbf{v} = (v_1, \cdots, v_d) \in \mathbb{R}^d$ and $R > 0$ be a positive real number. We denote by $\mathcal{S}_{\mathbf{v},R}^{(d)}$ the d-dimensional sphere centred around $\mathbf{v}$ of radius R. That is,

$$\mathcal{S}_{\mathbf{v},R}^{(d)} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{v} - \mathbf{x}\|_2 = R\}.$$

Furthermore, we denote by $\mathcal{B}_{\mathbf{v},R}^{(d)}$, the d-dimensional ball centred around $\mathbf{v}$ of radius R. That is,

$$\mathcal{B}_{\mathbf{v},R}^{(d)} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{v} - \mathbf{x}\|_2 \leq R\}.$$

For simplicity, if $\mathbf{v} = \mathbf{0} \in \mathbb{R}^d$, we write $\mathcal{S}_R^{(d)}$ and $\mathcal{B}_R^{(d)}$ respectively instead. Lemma 3 provides some observations on $\mathcal{S}_R^{(d)}$ and $\mathcal{B}_R^{(d)}$.

Lemma 3. *Let $R > 0$. Then*

1) *$A_R^{(d)} \triangleq \int\cdots\int_{\mathcal{S}_R^{(d)}} d\mathbf{x}$, which is the surface area of $\mathcal{S}_R^{(d)}$, has the following form*

$$A_R^{(d)} = \begin{cases} \frac{d\pi^{\frac{d}{2}} R^{d-1}}{\left(\frac{d}{2}\right)!}, & \text{if } d \text{ is even}, \\ 4\frac{\pi^{\frac{d-1}{2}} R^{d-1}}{\left(\frac{d-3}{2}\right)!} \prod_{\ell=1}^{\frac{d-3}{2}} \frac{2\ell}{2\ell+1}, & \text{if } d \text{ is odd} \end{cases}.$$

2) *$V_R^{(d)} \triangleq \int\cdots\int_{\mathcal{B}_R^{(d)}} d\mathbf{x}$, which is the volume of $\mathcal{B}_R^{(d)}$, has the following form*

$$V_R^{(d)} = \begin{cases} \frac{\pi^{\frac{d}{2}} R^d}{\left(\frac{d}{2}\right)!}, & \text{if } d \text{ is even}, \\ 4\frac{\pi^{\frac{d-1}{2}} R^d}{d\left(\frac{d-3}{2}\right)!} \prod_{\ell=1}^{\frac{d-3}{2}} \frac{2\ell}{2\ell+1}, & \text{if } d \text{ is odd} \end{cases}.$$

3) *$S_R^{(d)} \triangleq \int\cdots\int_{\mathcal{S}_R^{(d)}} e^{-k\|\mathbf{x}\|_2} d\mathbf{x}$, has the following form*

$$S_R^{(d)} = \begin{cases} \frac{d\pi^{\frac{d}{2}}}{\left(\frac{d}{2}\right)!} R^{d-1} e^{-kR}, & \text{if } d \text{ is even} \\ \frac{4\pi^{\frac{d-1}{2}}}{\left(\frac{d-3}{2}\right)!} \prod_{\ell=1}^{\frac{d-3}{2}} \frac{2\ell}{2\ell+1} R^{d-1} e^{-kR}, & \text{if } d \text{ is odd} \end{cases}.$$

4) *Let* $B_R^{(d)} \triangleq \int\cdots\int_{\mathcal{B}_R^{(d)}} e^{-k\|\mathbf{x}\|_2} d\mathbf{x}$. *Then there exists* $c_R \in (0, R)$ *such that*

$$B_R^{(d)} = \begin{cases} 2\frac{\pi^{\frac{d}{2}}}{(\frac{d}{2}-1)!} \cdot \frac{R^d}{d} e^{k(c_R - R)}, & \text{if } d \text{ is even} \\ 4\frac{\pi^{\frac{d-1}{2}}}{(\frac{d-3}{2})!}\varphi \cdot \frac{R^d}{d} e^{k(c_R-R)}, & \text{if } d \text{ is odd} \end{cases}$$

where $\varphi = \prod_{\ell=1}^{\frac{d-3}{2}} \frac{2\ell}{2\ell+1}$.

Proof. The proof can be found in Supplemental Material Section I. □

Using the observation made in Lemma 3, it is easy to verify the value of $\lambda_{\mathbf{v},L}$, which is given in Lemma 4.

Lemma 4. *Let* $L, k > 0$ *be two constants. Then* $\lambda_{\mathbf{v},L}$ *is independent of* $\mathbf{v}$, *and hence denoted by* λ_L. *We have,* $\lambda_L = \mu_L^{-1}$ *where*

$$\begin{aligned} \mu_L &= B_L^{(d)} + e^{-kL}\left((1+2L)^d - V_L^{(d)}\right) \\ &= 2\pi e^{-kL}\left[\frac{(d-1)!}{k^d}\left(e^{kL} - \sum_{i=0}^{d-1}\frac{(kL)^i}{i!}\right) - \frac{L^d}{d}\right] \\ &\quad \cdot \prod_{j=1}^{d-2} S_{3,d-1-j} + e^{-kL}(1+2L)^d, \text{ where} \end{aligned}$$

$$S_{3,j} = \int_0^\pi \sin^j x dx \text{ and}$$

$$\prod_{j=1}^{d-2} S_{3,d-1-j} = \begin{cases} \frac{\pi^{\frac{d}{2}-1}}{(\frac{d}{2}-1)!} & \text{if } d \text{ is even} \\ 2\frac{\pi^{\frac{d-3}{2}}}{(\frac{d-3}{2})!}\prod_{\ell=1}^{\frac{d-3}{2}}\frac{2\ell}{2\ell+1} & \text{if } d \text{ is odd} \end{cases}.$$

For the discussion in the following sections, we define

$$\begin{aligned} \mu_L^{(1)} &= e^{kL}\left(B_L^{(d)} - V_L^{(d)}\right) \\ &= 2\pi\left[\frac{(d-1)!}{k^d}\left(e^{kL} - \sum_{i=0}^{d-1}\frac{(kL)^i}{i!}\right) - \frac{L^d}{d}\right]\prod_{j=1}^{d-2} S_{3,d-1-j} \\ &= 2\pi\frac{(d-1)!}{k^d}\left(e^{kL} - \sum_{i=0}^{d}\frac{(kL)^i}{i!}\right)\prod_{j=1}^{d-2} S_{3,d-1-j} \geq 0 \end{aligned}$$

and

$$\mu_L^{(2)} = (1+2L)^d \geq 0.$$

Hence we have $\mu_L = e^{-kL}\left(\mu_L^{(1)} + \mu_L^{(2)}\right)$. In the following section, we discuss the algorithm to sample following the distribution with density function given in Eq. (2).

1) Sampling Based on BPM: Fix $\mathbf{v} \in \mathcal{D}$ to be the real data record. We note that the density function given can be divided to two cases depending on $\|\mathbf{x}-\mathbf{v}\|_2$. Let p_L be the probability that $\|\mathbf{x}-\mathbf{v}\| \leq L$. It is then easy to see that

$$p_L = \lambda_{\mathbf{v},L} B_L^{(d)} = \frac{B_L^{(d)}}{B_L^{(d)} + e^{-kL}\left((1+2L)^d - V_L^{(d)}\right)}.$$

In the following, we assume that such p_L has been calculated prior to the actual mechanism and is included as one of the inputs of the mechanism.

We can then rearrange the sampling of $f_{\mathbf{v}}^{(L)}$ to a two-step sampling process. First, we decide whether $\mathbf{x}$ is within L from $\mathbf{v}$ or $\mathbf{x}$ is farther than L away from $\mathbf{v}$. This step can be done by performing Bernoulli sampling with success probability p_L. Depending on the sampling result of this first step, it determines whether we sample $\mathbf{x} \in \mathcal{B}_{\mathbf{v},L}^{(d)}$ with distribution proportional to $e^{-k\|\mathbf{x}-\mathbf{v}\|}$, which happens with probability p_L or we sample $\mathbf{x} \in [-L, 1+L]^d \setminus \mathcal{B}_{\mathbf{v},L}^{(d)}$ with uniform distribution. Define $f_1^{(\mathbf{v},L)} : [-L, 1+L]^d \setminus \mathcal{B}_{\mathbf{v},L}^{(d)} \to \mathbb{R}_{\geq 0}$ and $f_2^{(\mathbf{v},L,k)} : \mathcal{B}_{\mathbf{v},L}^{(d)}$ such that for $\mathbf{x} \in [-L, 1+L]^d \setminus \mathcal{B}_{\mathbf{v},L}^{(d)}$ and $\mathbf{x}' \in \mathcal{B}_{\mathbf{v},L}^{(d)}$,

$$f_1^{(\mathbf{v},L)}(\mathbf{x}) = \frac{1}{(2+L)^d - V_L^{(d)}}, \text{ and}$$

$$f_2^{(\mathbf{v},L,k)}(\mathbf{x}') = \lambda_2 e^{-k\|\mathbf{v}-\mathbf{x}'\|_2}$$

where $\lambda_2 = \left(\int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},L}^{(d)}} e^{-k\|\mathbf{v}-\mathbf{x}\|_2} d\mathbf{x}\right)^{-1} = \left(B_L^{(d)}\right)^{-1}$. The full specification of the sampling algorithm is described in Algorithm 2.

Algorithm 2 BPM Sampling: $\mathbf{x} \leftarrow$ `BPMSampling`$(\mathbf{v}, k, L, p_L)$

Require: User's data record $\mathbf{v}$ of dimension d, decay parameter k, distance threshold L and the corresponding p_L
Ensure: Perturbed report $\mathbf{x}$ using BPM

```
Randomly sample a value p uniformly from [0, 1];
if p > p_L then
    Execute x ← Samplef1(v, L) to sample x uniformly across [−L, 1 + L]^d \ B^(d)_{v,L} with probability density function f_1^(v,L);
else
    Execute x ← Samplef2(v, k, L) to sample x with probability density function f_2^(v,L,k) in sample space B^(d)_{v,L};
end if
return the perturbed report x;
```

The following sections introduce the details of the sampling method from $f_1^{(\mathbf{v},L)}$ and $f_2^{(\mathbf{v},L,k)}$ through the protocols `Samplingf1` and `Samplingf2` respectively.

Sampling $\mathbf{x}$ *following* $f_1^{(\mathbf{v},L)}$. Since $f_1^{(\mathbf{v},L)}$ is a uniform distribution, we just need to randomly sample the data points in the domain while ensuring the distance between the real data record and randomized data record is beyond L. The full specification of the protocol `Samplingf1` is given in Algorithm 3.

As shown in Algorithm 3, we initialize the randomized data $\mathbf{x}$ as the real data record, then update each value in $\mathbf{x}$ by randomly sampling a value in $[-L, 1+L]$ and keep updating it until the distance between $\mathbf{x}$ and $\mathbf{v}$ is larger than L.

Sampling $\mathbf{x}$ *following* $f_2^{(\mathbf{v},L,k)}$. Recall that the sampling space of $f_2^{(\mathbf{v},L,k)}$ is $\mathcal{B}_{\mathbf{v},L}^{(d)}$, the d-dimensional ball centred around $\mathbf{v}$ with radius L. Using spherical substitution, we may express $f_2^{(\mathbf{v},L,k)}$ as a product of d independent marginal distribu-

Algorithm 3 Sampling $f_1^{(\mathbf{v},L)}$: $\mathbf{x} \leftarrow$ `Samplef1`$(\mathbf{v}, L)$

Require: User's data record $\mathbf{v}$ of dimension d and distance threshold L

Ensure: Perturbed report $\mathbf{x}$ following the distribution $f_1^{(\mathbf{v},L)}$

1: Set $\mathbf{x} = \mathbf{v}$;
2: **while** $\|\mathbf{x} - \mathbf{v}\|_2 \leq L$ **do**
3: **for** $i = 1, \cdots, d$ **do**
4: Uniformly sample $x_i \leftarrow [-L, 1 + L]$;
5: **end for**
6: Update $\mathbf{x} = (x_1, \cdots, x_d)$;
7: **end while**
8: **return** $\mathbf{x}$

tions on $r, \theta_1, \cdots, \theta_{d-1}$. Denote the marginal distributions of $f_2^{(\mathbf{v},L,k)}$ with respect to r by $f_2^{(\mathbf{v},L,k),(s,r)}(r)$. We have

$$\begin{aligned} f_2^{(\mathbf{v},L,k),(s,r)}(r) &= \lambda_2 r^{d-1} e^{-kr} \\ &\quad \cdot \prod_{i=1}^{d-2} \left(\int_0^{\pi} \sin^{d-1-i} d\theta_i \right) \cdot \int_0^{2\pi} d\theta_{d-1} \\ &= \lambda_{2,r} r^{d-1} e^{-kr} \end{aligned}$$

where

$$\lambda_{2,r} = \left(\int_0^L r^{d-1} e^{-kr} dr \right)^{-1} = \frac{k^d e^{kL}}{(d-1)!} \left(e^{kL} - \sum_{i=0}^{d-1} \frac{(kL)^i}{i!} \right)^{-1}.$$

Therefore, we can sample the radius r following $f_2^{(\mathbf{v},L,k),(s,r)}(r)$. Given the radius r, by design, any point $\mathbf{x}$ on the sphere $\mathcal{S}_{\mathbf{v},r}^{(d)}$ has the same probability density. Hence, once r is fixed, we may uniformly sample the point in $\mathcal{S}_{\mathbf{v},r}^{(d)}$ to obtain $\mathbf{x}$ that follows the density function $f_2^{(\mathbf{v},L,k),(s,r)}$. Such sampling can be achieved, for example, using the method proposed by Muller [14]. Algorithm 4 shows the details of sampling process following $f_2^{(\mathbf{v},L,k)}$.

Algorithm 4 Sampling $f_2^{(\mathbf{v},L,k)}$: $\mathbf{x} \leftarrow$ `Samplef2`$(\mathbf{v}, L, k)$

Require: User's data record $\mathbf{v}$ of dimension d, distance threshold L and decay parameter k

Ensure: Perturbed report $\mathbf{x}$ following the distribution $f_2^{(\mathbf{v},L,k)}$

1: Sample the radius r following the distribution $f_2^{(\mathbf{v},L,k),(s,r)}(r)$ with sample space $[0, L]$;
2: Sample a vector $\mathbf{t} = (t_1, \cdots, t_d)$ where $t_1 \sim \mathcal{N}(0, 1)$;
3: $\mathbf{x} = \mathbf{v} + r \cdot \mathbf{t}/||\mathbf{t}||_2$;
4: **return** $\mathbf{x}$;

We note that the first step of Algorithm 4 requires the sampling of the radius r following a distribution $f_2^{(\mathbf{v},L,k),(s,r)}(r)$ within $[0, L]$. Such step can be done by following the commonly used inverse sampling method. For completeness, we provide a brief discussion on this in Supplemental Material Section II.

2) Complexity analysis of the sampling algorithm: First we consider the complexities of the supporting subprotocols, `Samplef1` and `Samplef2`.

Proposition 5. *Let d and L be positive constants and $\mathbf{v} \in [0, 1]^d$. Suppose that $T_1(n)$ is the complexity of calculating the $L^{(2)}$ norm of a real-valued vector of length n and $T_2(a, b)$ is the complexity of uniformly sampling a real number from an interval $[a, b]$. Then the expected complexity of* `Samplef1` *is $(2 + 2L)T_1(1) + (1 + 2L)T_2(-L, 1 + L)$ if $d = 1$ and is at most $6T_1(d)+5dT_2(-L, 1+L)$ otherwise. In particular, if both $\|.\|_2$ and real uniform sampling can be done efficiently, then* `Samplef1` *can also be implemented efficiently in expectation.*

Proof. First, we define $N(d, L)$, the random variable representing the number of times Lines 3 up to 6 is repeated. It is easy to see that $N(d, L)$ is a geometric distribution with success probability p being equal to the probability that a uniformly sampled $\mathbf{x} \in [-L, 1+L]^d$ has a Euclidean distance of more than L from $\mathbf{v}$. It is easy to see that

$$p = 1 - \frac{V_L^{(d)}}{(1 + 2L)^d}.$$

Hence the expected value $\mathbb{E}(N(d, L)) = \frac{1}{p}$. We divide the case to three depending on the parity of d.

1) $d = 1$. Then the probability that a random real number in $[-L, 1 + L]$ not to be in $[v - L, v + L]$ is $\frac{1}{1+2L}$. Hence $N(d, L) = 1 + 2L$. Then when $d = 1$, the expected complexity of `Samplef1` is $(2L + 2)T_1(1) + (2L + 1)T_2(-L, 1 + L)$.
2) d is even. Let $d = 2t$ for some positive integer t. Then by Lemma 3, $V_L^{(d)} = \frac{\pi^t L^{2t}}{t!}$ and
$$N(2t, L) = \frac{1}{1 - \frac{\pi^t}{t!} \left(\frac{L}{1+2L} \right)^{2t}}.$$
It can be easily verified that $N(2t, L)$ is a decreasing function on t and when $N(2, L)$ is an increasing function on L. Hence $N(2t, L) \leq \lim_{L\to\infty} N(2, L) = \frac{1}{1-\frac{\pi}{4}} \leq 5$. This shows that $\mathbb{E}(N(d, L)) \leq 5$ when d is even.
3) $d \geq$ is odd. Let $d = 2t + 1$ for some positive integer t. Then by Lemma 3, we have $N(2t + 1, L)$ equals
$$\frac{1}{1 - \frac{4\pi^t}{(2t+1)(t-1)!} \prod_{\ell=1}^{t-1} \frac{2\ell}{2\ell+1} \left(\frac{L}{1+2L} \right)^{2t+1}}.$$
It can be easily verified that $N(2t+1, L)$ is again a decreasing function on t and $N(3, L)$ is an increasing function on L. Hence $N(2t + 1, L) \leq \lim_{L\to\infty} N(3, L) = \frac{1}{1-\frac{\pi}{9}} \leq 2 \leq 5$. This shows that $\mathbb{E}(N, d, L)) \leq 5$ when $d \geq 3$ is odd. Combined with the previous case, this shows that the expected complexity of `Samplef1` when $d \geq 2$ is at most $6T_1(d) + 5dT_2(-L, 1 + L)$ proving the claim.

□

Next, we consider the complexity of `Samplef2`. Since the claim can be easily verified, we state the result without proof.

Proposition 6. *Let d be a positive integer constant and L, k be positive real number constants. Furthermore, let $\mathbf{v} \in [0, 1]^d$. Suppose that the method by Muller [14] has*

complexity $T_3(L,k)$, the complexity of sampling from a standard normal distribution is T_4, the complexity of calculating $\|\cdot\|_2$ of a vector of length d is $T_1(d)$ and the complexity of computing a linear combination of two real-valued vector of length d be $T_5(d)$. Then the complexity of `Samplef2` *is $T_3(L,k)+dT_4+T_1(d)+T_5(d)$. In particular, if all such sub-protocols can be implemented efficiently, then* `Samplef2` *can also be implemented efficiently.*

Combining the claims from Propositions 5 and 6, we directly obtain the following result for the complexity of `BPMSampling`.

Theorem 7. *Let d, k and L be given constants and $\mathbf{v} \in [0,1]^d$. We use the same assumption for the complexities of the 5 sub-protocols used in both* `Samplef1` *and* `Samplef2` *which are provided in Propositions 5 and 6. Then:*

1) *If $d = 1$, the expected complexity of* `BPMSampling` *is at most $T_2(0,1)+\max\{(2+2L)T_1(1)+(1+2L)T_2(-L,1+L), T_3(L,k)+T_4+T_1(1)+T_5(1)\}$.*
2) *If $d \geq 2$, the expected complexity of* `BPMSampling` *is at most $T_2(0,1) + \max\{6T_1(d)+5dT_2(-L,1+L), T_3(L,k)+dT_4+T_1(d)+T_5(d)\}$.*

So in particular, if all 5 sub-protocols ($\|\cdot\|_2$, Uniform sampling, sampling based on Muller [14], standard normal sampling and linear combination calculation) can be implemented efficiently, then `BPMSampling` *can also be implemented efficiently in expectation.*

Remark 1. *In this work, we focus on the improvement of locally private K-means clustering algorithm under the d_χ-privacy definition. Because of this, our proposed bounded perturbation mechanism is described and analyzed in such setting. However, due to its generic form, it is easy to see that* `BPM` *is also applicable in various other data analytical tasks. Furthermore, as discussed in the next section, the increase of complexity with respect to the number of dimensions grows much more slowly compared to other traditional mechanisms such as Laplace mechanism. This suggests that* `BPM` *may provide better utility in applications with higher dimensional data compared to the use of traditional Laplace mechanism. Lastly, the perturbation mechanism* `BPM` *itself may be of independent theoretical interest in the study of differential privacy for multi-dimensional data.*

C. Privacy and Utility Analysis

In this section, we provide theoretical analysis on the privacy and utility of BPM as well as the privacy of Algorithm 1 when we use `BPM` as the $\epsilon\ d_E-$private mechanism $\mathcal{M}$.

1) Privacy Analysis: Theorem 8 shows that the proposed bounded perturbation mechanism satisfies ϵd_E-privacy guarantee for any $\epsilon > 0$.

Theorem 8. *Let $\epsilon > 0$ be the chosen privacy budget. The proposed bounded perturbation mechanism provides ϵd_E-privacy when k is set to be ϵ.*

Proof. Let $d_E(\mathbf{v},\mathbf{v}') \in \mathbb{R}_{>0}, \mathbf{x} \in \mathbb{R}^d$ and $\mathbf{v},\mathbf{v}' \in \mathcal{D}$ such that $\|\mathbf{v},\mathbf{v}'\|_2 = d_E(\mathbf{v},\mathbf{v}')$. Recall that we want to show that $\rho \triangleq \frac{f_{\mathbf{v}}^{(L)}(\mathbf{x})}{f_{\mathbf{v}'}^{(L)}(\mathbf{x})} \leq e^{\epsilon d_E(\mathbf{v},\mathbf{v}')}$. We remark that since $\epsilon d_E(\mathbf{v},\mathbf{v}') > 0$, we have $1 < e^{\epsilon d_E(\mathbf{v},\mathbf{v}')}$. We consider the ratio in four different cases depending on the relation between $\|\mathbf{x}-\mathbf{v}\|_2, \|\mathbf{x}-\mathbf{v}'\|_2$ and L.

1) Suppose that $L \geq \|\mathbf{x}-\mathbf{v}\|_2$ and $L \geq \|\mathbf{x}-\mathbf{v}\|_2$. Then
$$\rho = e^{k(\|\mathbf{x}-\mathbf{v}'\|_2-\|\mathbf{x}-\mathbf{v}\|_2)} \leq e^{k\|\mathbf{v}-\mathbf{v}'\|_2} = e^{\epsilon\cdot d_E(\mathbf{v},\mathbf{v}')}.$$
2) Suppose that $\|\mathbf{x}-\mathbf{v}\|_2 \leq L \leq \|\mathbf{x}-\mathbf{v}'\|_2$. Then
$$\begin{aligned}\rho &= e^{k(L-\|\mathbf{x}-\mathbf{v}\|_2)}\\ &\leq e^{k(\|\mathbf{x}-\mathbf{v}'\|_2-\|\mathbf{x}-\mathbf{v}\|_2)}\\ &\leq e^{k\|\mathbf{v}-\mathbf{v}'\|_2}\\ &= e^{\epsilon\cdot d_E(\mathbf{v},\mathbf{v}')}.\end{aligned}$$
3) Suppose that $\|\mathbf{x}-\mathbf{v}\|_2 \leq L \leq \|\mathbf{x}-\mathbf{v}'\|_2$. Then
$$\rho = e^{k(\|\mathbf{x}-\mathbf{v}'\|_2-L)} \leq 1 < e^{\epsilon\cdot d_E(\mathbf{v},\mathbf{v}')}.$$
4) Suppose that $L \leq \|\mathbf{x}-\mathbf{v}\|_2$ and $L \leq \|\mathbf{x}-\mathbf{v}'\|_2$. Then
$$\rho = e^{k(L-L)} = 1 < e^{\epsilon\cdot d_E(\mathbf{v},\mathbf{v}')}.$$

This completes the proof that if $k = \epsilon$, the mechanism described above provides the privacy guarantee of $\epsilon\ d_E$-privacy. □

Theorem 9 shows that using `BPM` in the proposed K-means clustering algorithm given in Algorithm 1, we have a private K-means clustering algorithm.

Theorem 9. *The proposed BPM-based K-means clustering algorithm satisfies $\epsilon\ d_E$-privacy.*

Proof. The proposed private K-means algorithm applies BPM to each user's report. Therefore, the user's randomized report that is sent to the server satisfies $\epsilon\ d_E$-privacy. After collecting reports from all the users, the server performs K-means clustering algorithm without any additional information from the user nor any further perturbation. Since any processing done by the server can be seen as post-processing that does not affect the privacy [12], [13], we can conclude that the proposed BPM-based K-means clustering algorithm also satisfies ϵd_E-privacy. □

2) Utility Analysis: We first analyze the property of the perturbation mechanism in terms of *Expectation* and *Variance* followed by the analysis of the distance property provided by BPM for K-means clustering algorithm.

First, we recall the notations. We let $k > 0$ be the decay parameter and $L > 0$ be the distance threshold. Furthermore, we fix $\mathbf{v} \in \mathcal{D}$ to be a user's real data. In this section, we investigate the distribution of the random variable $\mathbf{X} \sim f_{\mathbf{v}}^{(L)}$ which represents the output of the protocol `BPMSampling`$(\mathbf{v},k,L,p_L)$. We denote by $\mathbf{e}_{\mathbf{v},L} = \mathbb{E}(\mathbf{X}) \in \mathcal{R}_L$ and $V = \mathrm{Var}(\mathbf{X}) \in \mathbb{R}^{d\times d}$.

Lemma 10. *The expected value of the random variable $\mathbf{X}$ is*
$$\mathbf{e} = \mathbf{v} + (1+2L)^d\lambda_L e^{-kL}\left(\frac{1}{2}-v_1,\cdots,\frac{1}{2}-v_d\right).$$

Proof. Let $i = 1, \cdots, d$. Consider the expected value of X_i, defined as $\mathbb{E}(X_i) = \int\cdots\int_{\mathbf{x}\in\mathcal{R}_L} x_i f_{\mathbf{v}}^{(L)}(\mathbf{x})d\mathbf{x}$. Let $\Delta_i = \mathbb{E}(X_i) - v_i$. Hence $\Delta_i = \int\cdots\int_{\mathbf{x}\in\mathcal{R}_L}(x_i - v_i)f_{\mathbf{v}}^{(L)}(\mathbf{x})d\mathbf{x}$. Define $\Delta_{i,1} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},\mathbf{L}}^{(d)}}(x_i - v_i)\lambda_L e^{-k\|\mathbf{x}-\mathbf{v}\|_2}d\mathbf{x}$, the part of Δ_i where $\|\mathbf{x} - \mathbf{v}\|_2 \leq L$. Next, we also define $\Delta_{i,2} = \int\cdots\int_{\mathbf{x}\in\mathcal{R}_L}(x_i - v_i)\lambda_L e^{-kL}d\mathbf{x}$ and $\Delta_{i,3} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},\mathbf{L}}^{(d)}}(x_i - v_i)\lambda_L e^{-kL}d\mathbf{x}$. Then we have $\Delta_i = \Delta_{i,1} + \Delta_{i,2} - \Delta_{i,3}$.

1) First we consider $\Delta_{i,1}$. Intuitively, this is simply a bounded multi-dimensional variant of Laplace distribution with a ball of finite radius centred at $\mathbf{v}$. So we can expect $\Delta_{i,1}$ to be 0. The claim below formally proves this intuition.

Claim 11. $\Delta_{i,1} = 0$.

Proof. First, we consider the case when $i \leq d-2$. Then $x_i - c_i = r\sin\theta_1 \cdot \ldots \cdot \sin\theta_{i-1} \cdot \cos\theta_i$. Using the multidimensional spherical representation, we have

$$\begin{aligned}
0 \leq \Delta_i &= \lambda\left(\int_0^L r^d e^{-kr}dr\right) \\
&\cdot \prod_{j=1}^{i-1}\left(\int_0^\pi \sin^{d-j}\theta_j\right)\cdot\int_0^\pi \sin^{d-1-i}\theta_i\cos\theta_i d\theta_i \\
&\cdot \prod_{j=i+1}^{d-2}\left(\int_0^\pi \sin^{d-1-j}\theta_j d\theta_j\right)\int_0^{2\pi} d\theta_{d-1} \\
&\leq \lambda L^{d+1}\cdot\prod_{j=1}^{i-1}\pi\cdot\prod_{j=i+1}^{d-2}\pi \\
&\cdot 2\pi\cdot\int_0^\pi \sin^{d-1-i}\theta_i\cos\theta_i d\theta_i
\end{aligned}$$

where the last inequality is based on the observations that for $0 \leq r \leq L, r^d e^{-kr} \leq L^d$ and $-1 \leq \sin\theta \leq 1$ for any θ. Furthermore, we note that

$$\int_0^\pi \sin^{d-1-i}\theta_i\cos\theta_i d\theta_i = \left(\frac{1}{d-i}\sin^{d-i}\theta_i\Big|_0^\pi\right) = 0.$$

Hence $\Delta_{i,1} = 0$.
Next we consider the case when $i = d-1, d$. It is easy to see that

$$\begin{aligned}
0 \leq \Delta_{i,1} &= \lambda\left(\int_0^L r^d e^{-kr}dr\right) \\
&\cdot\prod_{j=1}^{d-2}\left(\int_0^\pi \sin^{d-j}\theta_j\right) \\
&\cdot\int_0^{2\pi} g(\theta_{d-1})d\theta_{d-1} \\
&\leq \lambda L^{d+1}\pi^{d-2}\int_0^{2\pi} g(\theta_{d-1})d\theta_{d-1}
\end{aligned}$$

where $g(\theta_{d-1}) = \cos\theta_{d-1}$ when $i = d-1$ and $g(\theta_{d-1}) = \sin\theta_{d-1}$ when $i = d$. Note that $\int_0^{2\pi} g(\theta_{d-1})d\theta_{d-1} = 0$. Hence we again have $\Delta_{i,1} = 0$ for $i = d-1, d$. □

2) Next, we consider $\Delta_{i,2}$. It is easy to see that

$$\begin{aligned}
\Delta_{i,2} &= \lambda_L e^{-kL}\int\cdots\int_{\mathbf{x}\in\mathcal{R}_L}(x_i - v_i)d\mathbf{x} \\
&= (1+2L)^d\lambda_L e^{-kL}\left(\frac{1}{2} - v_i\right).
\end{aligned}$$

3) Lastly, we consider $\Delta_{i,3}$. Then $\Delta_{i,3} = \lambda_L e^{-kL}\int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},L}^{(d)}}(x_i - v_i)d\mathbf{x} = 0$ where we again use a similar proof as Claim 11.

Then we have $\Delta_i = (1+2L)^d\lambda_L e^{-kL}\left(\frac{1}{2} - v_i\right)$. □

Lemma 10 confirms that the random variable $\mathbf{X}$ following the distribution with density function $f_{\mathbf{v}}^{(L)}$ is not an unbiased estimator of the original data record $\mathbf{v}$. However, as L increases, it is easy to see that the bias decreases and as L approaches ∞, the bias approaches 0 confirming that our BPM approaches the standard unbounded multi-dimensional Laplace mechanism as we allow L grows to ∞.

Lemma 12. *Let $V = \mathrm{Var}(\mathbf{X}) \in \mathbb{R}^{d\times d}$ be the variance of the random variable $\mathbf{X}$ and $v_{i,j}$ be the i-th row and j-th column of V. For $i, j = 1, \cdots, d$, we have*

1) *If $i \neq j$,*

$$|v_{i,j}| \leq \frac{(1+2L)^d}{4}\lambda_L e^{-kL}\left|1 - (1+2L)^d\lambda_L e^{-kL}\right|,$$

2) *If $i = j$, there exists $0 < C_L^{(d+1)} < C_L^{(d-1)} < L$ such that*

$$v_{i,i} \leq \left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}} - 1}{e^{kc_L^{(d-1)}} - 1}\right).$$

Proof. Let $i, j = 1, \cdots, d$ and

$$v_{i,j} = \mathbb{E}\left((X_i - \mathbb{E}(X_i))\,(X_j - \mathbb{E}(X_j))\right)$$

be the i-th row and j-th column of V. Then we have

$$\begin{aligned}
v_{i,j} &= \mathbb{E}\left((X_i - e_i)\,(X_j - e_j)\right) = \mathbb{E}(X_i X_j) - e_i e_j \\
&= \mathbb{E}((X_i - v_i)(X_j - v_j)) - (e_i - v_i)(e_j - v_j).
\end{aligned}$$

Now we consider $\Gamma_{i,j} \triangleq \mathbb{E}((X_i - v_i)(X_j - v_j))$. First, we consider the case when $i < j$. Define $\Gamma_{i,j,1} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},L}^{(d)}}(x_i - v_i)(x_j - v_j)\lambda_L e^{-k\|\mathbf{x}-\mathbf{v}\|_2}d\mathbf{x}$, $\Gamma_{i,j,2} = \int\cdots\int_{\mathbf{x}\in\mathcal{R}_L}(x_i - v_i)(x_j - v_j)\lambda_L e^{-kL}d\mathbf{x}$ and $\Gamma_{i,j,3} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}_{\mathbf{v},L}^{(d)}}(x_i - v_i)(x_j - v_j)\lambda_L e^{-kL}d\mathbf{x}$.

1) We consider $\Gamma_{i,j,1}$ and $\Gamma_{i,j,3}$. In both cases, $\Gamma_{i,j,1} = \Gamma_{i,j,3} = 0$ which can be shown by using a simple modification of the proof of Claim 11.
2) We consider $\Gamma_{i,j,2}$. Here we have

$$\begin{aligned}
\Gamma_{i,j,2} &= \lambda_L e^{-kL}(1+2L)^{d-2}\int_{-L}^{1+L}(x_i - v_i)dx_i \\
&\cdot\int_{-L}^{1+L}(x_j - v_j)dx_j \\
&= (1+2L)^d\lambda_L e^{-kL}\left(\frac{1}{2} - v_i\right)\left(\frac{1}{2} - v_j\right).
\end{aligned}$$

So when $i \neq j$,

$$\begin{aligned}|v_{i,j}| &= \left|(1+2L)^d\lambda_L e^{-kL}\left(\frac{1}{2}-v_i\right)\left(\frac{1}{2}-v_j\right)\right. \\ &\quad \left.-(1+2L)^{2d}\lambda_L^2 e^{-2kL}\left(\frac{1}{2}-v_i\right)\left(\frac{1}{2}-v_j\right)\right| \\ &\leq \frac{(1+2L)^d}{4}\lambda_L e^{-kL}\left|1-(1+2L)^d\lambda_L e^{-kL}\right|\end{aligned}$$

Now we consider the case when $i = j$. Let $\Gamma_i \triangleq \mathbb{E}((X_i - v_i)^2)$. Define $\Gamma_{i,1} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}^{(d)}_{\mathbf{v},L}}(x_i - v_i)^2\lambda_L e^{-k\|\mathbf{x}-\mathbf{v}\|_2}d\mathbf{x}$, $\Gamma_{i,2} = \int\cdots\int_{\mathbf{x}\in\mathcal{R}_L}(x_i - v_i)^2\lambda_L e^{-kL}d\mathbf{x}$ and $\Gamma_{i,3} = \int\cdots\int_{\mathbf{x}\in\mathcal{B}^{(d)}_{\mathbf{v},L}}(x_i - v_i)^2\lambda_L e^{-kL}d\mathbf{x}$.

1) Note that the calculation of $\Gamma_{i,1}$ is similar to the calculation of $v_{i,i}$ in Claim 11 where $\lambda_L\int_0^L r^{d-1}e^{kr}dr$ is replaced by $\lambda_L\int_0^L r^{d+1}e^{-kr}dr$. Hence we also divide the calculation based on the value of i.
   - $i = 1,\cdots,d-2$. Using the same method as that of Claim 11, we have that when $i = 1,\cdots,d-2$,

$$\begin{aligned}\Gamma_{i,1} &= 2\pi\lambda_L e^{-kL}\frac{(d+1)!}{k^{d+2}} \\ &\quad\cdot\left(e^{kL}-\sum_{i=0}^{d+1}\frac{(kL)^i}{i!}\right)\cdot\left(\prod_{j=1}^{i-1}S_{3,d+1-j}\right) \\ &\quad\cdot(S_{3,d-1-i}-S_{3,d+1-i})\left(\prod_{j=i+1}^{d-2}S_{3,d-1-j}\right)\end{aligned}$$

     where $S_{3,i}$ is as defined in the proof of Lemma 3 which can be found in Supplemental Material I.
   - $i = d-1, d$. Then using the same method as that of Claim 11,

$$\Gamma_{i,1} = \pi\lambda_L\frac{(d+1)!}{k^{d+2}e^{kL}}\left(e^{kL}-\sum_{i=0}^{d+1}\frac{(kL)^i}{i!}\right)\prod_{j=1}^{d-2}S_{3,d+1-j}.$$

2) Next, we consider $\Gamma_{i,2}$. We have

$$\Gamma_{i,2} = (1+2L)^d\lambda_L e^{-kL}\left(\frac{1+L+L^2}{3}-v_i+v_i^2\right).$$

3) Lastly, we consider $\Gamma_{i,3}$. We will again divide the analysis based on the value of i.
   - $i = 1,\cdots,d-2$. Using the same method as before, when $i = 1,\cdots,d-2$,

$$\begin{aligned}\Gamma_{i,3} &= 2\pi\lambda_L e^{-kL}\frac{L^{d+2}}{d+2}\left(\prod_{j=1}^{i-1}S_{3,d+1-j}\right) \\ &\quad\cdot(S_{3,d-1-i}-S_{3,d+1-i})\left(\prod_{j=i+1}^{d-2}S_{3,d-1-j}\right).\end{aligned}$$

   - $i = d-1, d$. Then using the same method as before, when $i = d-1$ or $i = d$,

$$\Gamma_{i,3} = \pi\lambda_L e^{-kL}\frac{L^{d+2}}{d+2}\prod_{j=1}^{d-2}S_{3,d+1-j}.$$

Combining the analysis above, we have the following result on Γ_i.

1) If $i = 1,\cdots,d-2$, there exists $c_L^{(d+1)} \in \left(0, c_L^{(d-1)}\right)$ such that

$$\Gamma_i = \frac{\Gamma_i^{(1)}+\Gamma_i^{(2)}}{\mu_L^{(1)}+\mu_L^{(2)}}$$

where

$$\begin{aligned}\Gamma_i^{(1)} &= 2\pi\left(\prod_{j=1}^{i-1}S_{3,d+1-j}\right)(S_{3,d-1-i}-S_{3,d+1-i}) \\ &\quad\cdot\left(\prod_{j=i+1}^{d-2}S_{3,d-1-j}\right)\frac{L^{d+2}}{d+2}\left(e^{kc_L^{(d+1)}}-1\right)\end{aligned}$$

and

$$\Gamma_i^{(2)} = (1+2L)^d\left(\frac{1+L+L^2}{3}-v_i+v_i^2\right).$$

Note that for any $a,b,c,d > 0$, $\frac{a+b}{c+d} \leq \frac{a}{b}+\frac{c}{d}$. Hence $\Gamma_i \leq \frac{\Gamma_i^{(1)}}{\mu_L^{(1)}}+\frac{\Gamma_i^{(2)}}{\mu_L^{(2)}}$. It is easy to see that $\frac{\Gamma_i^{(2)}}{\mu_L^{(2)}} = \left(\frac{1+L+L^2}{3}-v_i+v_i^2\right)$. On the other hand,

$$\begin{aligned}\frac{\Gamma_i^{(1)}}{\mu_L^{(1)}} &= L^2\frac{d}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right) \\ &\quad\cdot\prod_{j=1}^{i-1}\frac{S_{3,d+1-j}}{S_{3,d-1-j}}\cdot\frac{S_{3,d-1-i}-S_{3,d+1-i}}{S_{3,d-1-i}} \\ &= \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right).\end{aligned}$$

Hence when $i = 1,\cdots,d-2$, we have the following upper bound

$$\Gamma_i \leq \left(\frac{1+L+L^2}{3}-v_i+v_i^2\right)+\frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right).$$

2) On the other hand, if $i = d-1$ or $i = d$,

$$\Gamma_i = \lambda_L e^{-kL}\left(\Gamma_i^{(1)}+\Gamma_i^{(2)}\right)$$

where

$$\Gamma_i^{(1)} = \pi\left(\prod_{j=1}^{d-2}S_{3,d+1-j}\right)\frac{L^{d+2}}{d+2}\left(e^{kc_L^{(d+1)}}-1\right)$$

and

$$\Gamma_i^{(2)} = (1+2L)^d\left(\frac{1+L+L^2}{3}-v_i+v_i^2\right).$$

Using the same upper bound as the previous case, we have $\frac{\Gamma_i^{(2)}}{\mu_L^{(2)}} = \frac{1+L+L^2}{3}-v_i+v_i^2$ and

$$\begin{aligned}\frac{\Gamma_i^{(1)}}{\mu_L^{(1)}} &= \frac{1}{2}L^2\frac{d}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right)\prod_{j=1}^{d-2}\frac{S_{3,d+1-j}}{S_{3,d-1-j}} \\ &= \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right)\end{aligned}$$

providing the same bound for Γ_i.

So for any $i = 1, \cdots, d$, we have

$$\Gamma_i \leq \left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right). \tag{3}$$

Then $v_{i,i} = \Gamma_i - (e_i - v_i)^2 \leq \Gamma_i$, which implies

$$v_{i,i} \leq \left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right)$$

as claimed. □

Despite its importance in the study of random variable and its distribution, in general, the variance of a multi-dimensional random variable is a matrix which is not equipped with a total ordering. So, to facilitate any theoretical accuracy comparison, considering that $\mathbf{X}$ is designed to be an estimator of the value $\mathbf{v}$, we provide the analysis of the random variable representing the distance $\|\mathbf{X} - \mathbf{v}\|_2$ in Lemma 13.

Lemma 13. *Let $\mathbf{v} \in \mathcal{D}$ and $\mathbf{X} \in \mathcal{R}_L$ follows the distribution $f_{\mathbf{v}}^{(L)}$ with variance matrix $V \in \mathbb{R}^{d\times d}$ and for $i = 1, \cdots, d$, we denote $\Gamma_i \triangleq \mathbb{E}((X_i - v_i)^2)$. Then*

$$\mathbb{E}(\|\mathbf{X} - \mathbf{v}\|_2^2) = \sum_{i=1}^{d} \Gamma_i.$$

Hence, there exists $0 < c_L^{(d+1)} \leq c_L^{(d-1)} < L$ such that

$$\mathbb{E}(\|\mathbf{X} - \mathbf{v}\|_2^2) \leq d\left(\frac{1+L+L^2}{3}\right) + \left(\sum_{i=1}^{d} v_i^2 - v_i\right) + \frac{dL^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right).$$

Proof. First, we consider the expected value of $\|\mathbf{v} - \mathbf{X}\|_2^2$.

$$\mathbb{E}(\|\mathbf{X} - \mathbf{v}\|_2^2) = \sum_{i=1}^{d} \mathbb{E}((x_i - v_i)^2) = \sum_{i=1}^{d} \Gamma_i.$$

Since

$$\Gamma_i \leq \left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{L^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right),$$

as shown in Eq. (3), we obtain the claimed upper bound. □

Remark 2. *Here we briefly discuss how to choose L. One way to optimize L is to find L that minimizes the average value ζ of $\mathbb{E}\left(\|\mathbf{X} - \mathbf{v}\|_2^2\right)$, the expected noise, for $\mathbf{v} \in [0,1]^d$, that is*

$$\zeta = \frac{\int_{\mathbf{v}\in[0,1]^d} \mathbb{E}\left(\|\mathbf{X} - \mathbf{v}\|_2^2\right) d\mathbf{v}}{\int_{\mathbf{v}\in[0,1]^d} d\mathbf{v}}.$$

Note that by definition, $\mathbb{E}\left(\|\mathbf{X} - \mathbf{v}\|_2^2\right) = C + (1+2L)^d \lambda_L e^{-kL} \sum_{i=1}^{d}(v_i^2 - v_i)$ for some value C that is independent of $\mathbf{v}$. Hence $\zeta = C + (1+2L)^d \lambda_L e^{-kL} \sum_{i=1}^{d} \int_0^1 x^2 - xdx = C - \frac{d}{6}(1+2L)^d \lambda_L e^{-kL}$. Note that ζ is a function of L and it is independent of the private value $\mathbf{v}$. Hence, to optimize the choice of L, we may use gradient descent or any other numerical techniques to estimate the smallest value of L with the desired value of ζ.

Suppose that we have K groups of users where the i-th group has n_i users, $U_{i,1}, \cdots, , U_{i,n_i}$. For any $i = 1, \cdots, K$ and $j = 1, \cdots, n_i,$, suppose that user $U_{i,j}$ holds $\mathbf{v}^{(i,j)} \in \mathcal{D}$. For each $i = 1, \cdots, K$, we define $\mathbf{v}^{(i)} \in \mathcal{D}$, the true central of the i-th group, then, $\mathbf{v}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{v}^{(i,j)}$. Lemma 14 provides an analysis of the distance property between the real centroids and perturbed centroids provided by the proposed method.

Lemma 14. *For $i = 1, \cdots, K$ and $j = 1, \cdots, n_i$, suppose that the report of $U_{i,j}$ is $\mathbf{x}^{(i,j)}$ and $\mathbf{x}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}^{(i,j)}$. Then $\mathbb{E}(\mathbf{x}^{(i)}) = \mathbf{v}^{(i)} + (1+2L)^d \lambda_L e^{-kL}\left(\mathbf{h} - \mathbf{v}^{(i)}\right)$ where $\mathbf{h} = \left(\frac{1}{2}, \cdots, \frac{1}{2}\right) \in \mathbb{R}^d$. Furthermore, letting $E = \mathbb{E}(\|\mathbf{x}^{(i)} - \mathbf{v}^{(i)}\|_2^2)$, we have*

$$E \leq \frac{1}{n_i} V + 2^{-(2d-1)} \frac{n_i - 1}{n_i} (1+2L)^{2d} \lambda_L^2 e^{-2kL}$$

where $V = d\left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{dL^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right)$.

Proof. First, we consider $\mathbb{E}(\mathbf{x}^{(i)})$. We have

$$\begin{aligned}\mathbb{E}(\mathbf{x}^{(i)}) &= \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbb{E}(\mathbf{x}^{(i,j)}) \\ &= \mathbf{v}^{(i)} + (1+2L)^d \lambda_L e^{-kL}\left(\mathbf{h} - \mathbf{v}^{(i)}\right).\end{aligned}$$

Next, we consider E.

$$\begin{aligned}E &= \mathbb{E}\left(\|\mathbf{x}^{(i)} - \mathbf{v}^{(i)}\|_2^2\right) \\ &= \frac{1}{n_i^2}\sum_{w=1}^{d} \mathbb{E}\left(\left(\sum_{j=1}^{n_i}(x_w^{(i,j)} - v_w^{(i,j)})\right)^2\right) \\ &= \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathbb{E}\left(\|\mathbf{x}^{(i,j)} - \mathbf{v}^{(i,j)}\|_2^2\right) \\ &\quad + \frac{2}{n_i^2}\sum_{u=1}^{n_i}\sum_{u'=1}^{u-1}\left((1+2L)^{2d}\lambda_L^2 e^{-2kL}\left(\mathbf{h} - \mathbf{v}^{(i,u)}\right)\right. \\ &\quad \left.\cdot\left(\mathbf{h} - \mathbf{v}^{(i,u')}\right)^T\right).\end{aligned}$$

So we have

$$E \leq \frac{1}{n_i} V + 2^{-(2d)} \frac{n_i - 1}{n_i} (1+2L)^{2d} \lambda_L^2 e^{-2kL}$$

where $V = d\left(\frac{1+L+L^2}{3} - v_i + v_i^2\right) + \frac{dL^2}{d+2}\left(\frac{e^{kc_L^{(d+1)}}-1}{e^{kc_L^{(d-1)}}-1}\right)$, as claimed. □

This shows that assuming the K-means clustering algorithm that is used by the server in Line 7 of Algorithm 1 outputs the same clusters as the one generated in the clear, the compactness of each cluster can be upper bounded by some value that depends on the compactness of the original clusters.

IV. Experimental Evaluation

A. Experimental Settings

1) Dataset: We implement the proposed method, which we will denote by DKdP for short, over three different real datasets, all of which can be found on the UCI machine learning repository used for the clustering problem.

- Seeds: seeds dataset consists of 210 data points of three wheat types. Each data point has 7 geometric attributes of wheat kernels.
- Travel Review: this data set is populated by data obtained from crawling the website TripAdvisor.com. Including 980 reviews on destinations in 10 categories across East Asia. We simply classify the reviewers into 10 clusters.
- 3D Road Network: this data set was constructed by adding elevation information to a 2D road network in North Jutland, Denmark. We extract 500 instances from area of $[9.5, 10] \times [56.6, 56.8] \times [20, 40]$. In addition, we extract $1,000$, $10,000$, and $100,000$ instances respectively from the whole area randomly to show the scalability of the proposed method and the effect of the population size.

2) Metric: K-means clustering is conducted with respect to Euclidean distance by minimizing Eq. 1. Therefore, we utilize a typical metric used for K-means clustering quality evaluation, *Sum of Squared Error (SSE)* , to evaluate the effect of perturbation on the compactness of the clusters. Specifically,

$$SSE = \sum_{i=1}^{K} \sum_{\mathbf{v}_j \in \mathbb{C}_i} \|\mathbf{v}_j - c_i\|^2$$

where c_i refers to the centroid of the cluster $\mathbb{C}_i$ and it is calculated as the mean value of the reports obtained from the members of the cluster $\mathbb{C}_i$.

Furthermore, we utilize *Relative Error (RE)* defined as follow to measure the distance between the centroid calculated based on the noisy reports and real reports.

$$RE = \sum_{i=1}^{K} \frac{1}{|\mathbb{C}_i|} \left\| \sum_{\mathbf{v}_j \in \mathbb{C}_i} \mathbf{v}_j - \sum_{\mathbf{x}_j \in \hat{\mathbb{C}}_i} \mathbf{x}_j \right\|_2 .$$

3) Comparison: Since there is no d_χ-private mechanism for K-means clustering problem in the literature for a fair comparison, in order to provide the baseline performance of privacy-preserving K-means clustering algorithm to be compared to our proposed mechanism, we utilize a standard solution using Laplace mechanism to provide local differential privacy to each dimension of the private data. In such case, the privacy budget is equally divided to provide local differential privacy to each dimension independently. However, we would like to note that such comparison may not be fair since the baseline is using a different differential privacy guarantee. For a better comparison, we may consider an alternative mechanism obtained by modifying the traditional Laplace mechanism to satisfy the same privacy guarantee as the one claimed in BPM. However, since such mechanism is not a common mechanism and due to the space limitation, we have provided such comparison and discussion in the Supplemental Material Section III.

In our experiments, we are interested to compare the methods with respect to various parameters where each experimental result is obtained by taking the average of 50 independent experiments measuring the same aspect (the box plot shows all 50 results for each setting).

B. Experimental Results

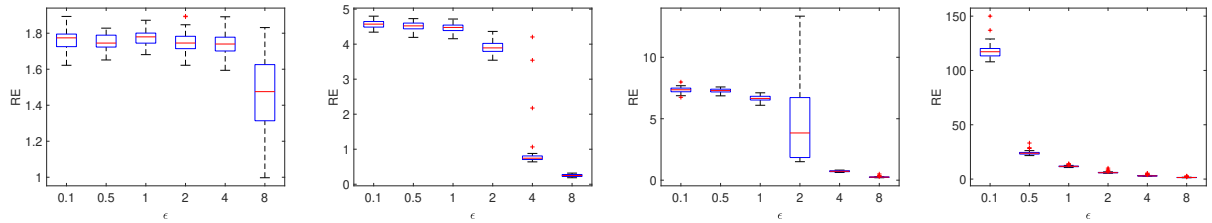
1) Performance in terms of Relative Error: First, we present the performance comparison based on the metric *RE*. In our experiment, we measure the average relative error for various values of $\epsilon \in \{0.1, 0.5, 1, 2, 4, 8\}$ in three different datasets we have discussed before. In these three datasets, we set the value of L to be $2, 6$, and 10 respectively.

First, we consider the effect of the change of ϵ to the metric *Relative Error*. As can be observed in Figures 1, 2, and 3, as ϵ increases, the relative error decreases. This is supported by the theoretical analysis, as can be observed in Lemmas 10 and 13. More specifically, we see that both the expected bias and expected distance of the perturbed report with the original value is a decreasing function of k, which is set to be ϵ to provide ϵd_E-differential privacy guarantee. Hence as ϵ increases, both values decreases, providing reports with statistically smaller noise, which in turn decreases the relative error. This can be observed, for example, in Fig. 1b, which has the relative error to decrease from over 4 when $\epsilon \leq 1$ to approximately 0.25 when ϵ reaches 8. Similar trend can be observed in other settings.

Despite similar trend being observed in other datasets, it can be observed that the gap is found to be much larger when considering the Review dataset, as observed in Fig. 2. More specifically, for $\epsilon = 0.1$, the relative error figure for the baseline is up to 250, which is much higher. This can be attributed to the higher dimension of the data considered in such dataset. More specifically, the baseline figure is obtained by performing traditional Laplace mechanism to perturb each dimension of the data independently. Due to this independent treatment to each dimension, the privacy budget needs to be divided equally to perturb each dimension, which is not required in our protocol. Note that this means that for each dimension, given the privacy budget ϵ, each dimension has privacy budget $\frac{\epsilon}{d}$. It is well known that for Laplace mechanism for data with sensitivity 1 and privacy budget $\frac{\epsilon}{d}$, the expected squared distance, or equivalently, its statistical variance, is $O(d^2)$. Now, combining the d dimensions, we have that the overall expected squared distance between the perturbed report and the original data is $O(d^3)$. On the other hand, as observed in Lemma 13, the expected squared distance of our perturbation method is $O(d)$. It is hence easy to see that as d increases, we can expect that this gap to increase.

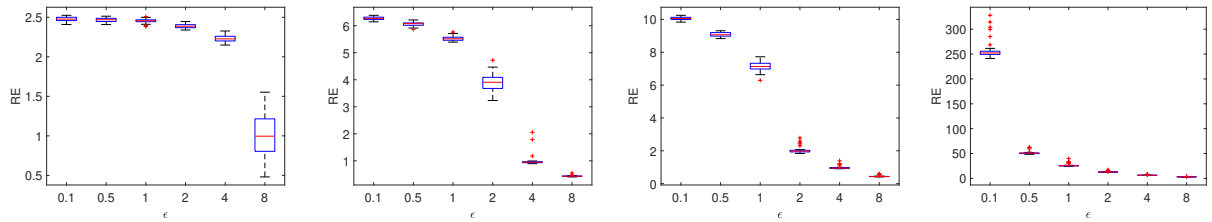
2) Performance in terms of SSE: In this section, we examine the performance of the proposed method in terms of *SSE* over three datasets. As has been done in the previous section, we vary the privacy budget ϵ and bound L to compare the accuracy of DKdP and Baseline.

Fig. 4 provides the SSE for both DKdP and the Baseline for various ϵ and L using datasets Seeds, Review and Network respectively. We note that in all cases, as ϵ increases DKdP



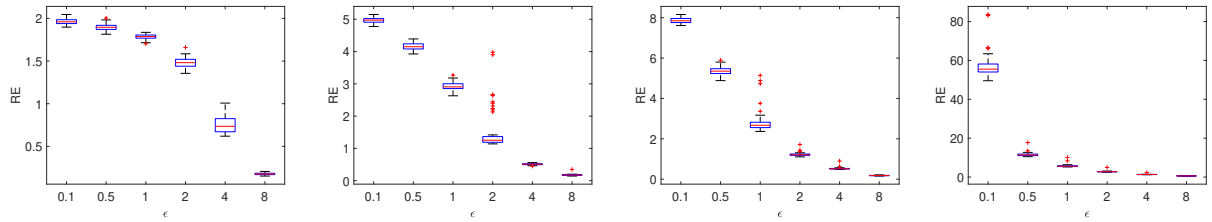
(a) DKdP: $L = 2$ (b) DKdP: $L = 6$ (c) DKdP: $L = 10$ (d) Baseline

Fig. 1. Performance by varying ϵ in terms of *RE* over Seeds dataset.



(a) DKdP: $L = 2$ (b) DKdP: $L = 6$ (c) DKdP: $L = 10$ (d) Baseline

Fig. 2. Performance by varying ϵ in terms of *RE* over Travel Review dataset.



(a) DKdP: $L = 2$ (b) DKdP: $L = 6$ (c) DKdP: $L = 10$ (d) Baseline

Fig. 3. Performance by varying ϵ in terms of *RE* over 3D Road Network dataset.

exhibits a much observable improvement in SSE compared to the Baseline. This is in fact more noticeable in the Review dataset which has data with larger dimensions, supporting our argument that DKdP and BPM that we have proposed possess higher advantage compared to traditional solutions when the dimension is relatively large.

3) Effect of the population size: To evaluate the effect of the population size, we randomly sample $1,000$, $10,000$, and $100,000$ samples from Network dataset and measure the average *SSE* and *RE* values. We fixed $L = 10$ and vary the privacy budget ϵ to 0.1, 1, and 4.

Fig. 5 shows the results. As expected, the increase in population provides us with a more accurate estimate, which in turns provides a more accurate clustering. This effect can be observed from the figures where the increase in the population size causes both averages to also decrease accordingly.

V. Related work

Differentially private K-means clustering has been extensively studied in the literature. However, the majority of the works focus on the centralized setting where a trusted data curator that has access to all users' real data records. The first version of differentially private K-means clustering algorithm was proposed by Blum *et al.* [15] in the interactive setting. In their scheme, the Laplacian noise is added to each iteration step which necessitates the splitting of privacy budget. To improve the utility of clustering, the works [16], [17] proposed optimal privacy budget allocation method while Ni *et al.* [18] proposed the use of cluster merging as well as adaptive noise to perturb centroids in each iterative computation. Mohan *et al.* [19] proposed a differentially private analysis system that relies on the sample-and-aggregate framework [20] which samples data points randomly to different buckets, where the local K-means algorithm is performed. The overall clustering result is then obtained by taking the average from the different buckets. Zhang *et al.* [21] proposed a generic algorithm-based private K-means clustering, which chooses the candidates for the next iteration through exponential mechanism. Lu and Shen [22], [23] addressed the non-convergence problem in the existing differentially private K-means algorithms. Specifically, they proposed a method which controls the orientation of the centroids movement over the iterations to ensure the convergence. This is done by injecting DP noise in a selected area. In [16], [24], a non-interactive setting is considered

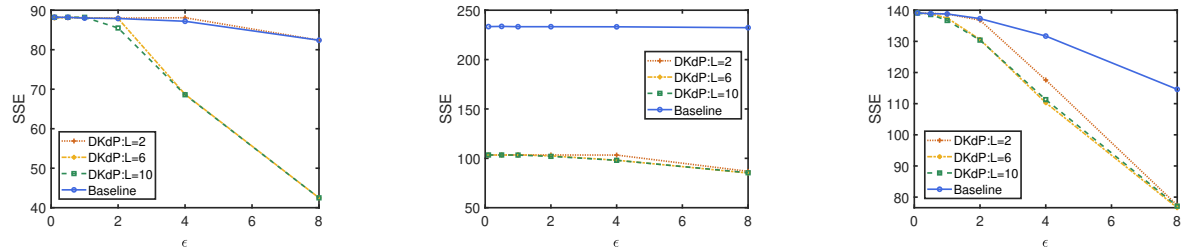


Fig. 4. Performance by varying ϵ in terms of *SSE*.

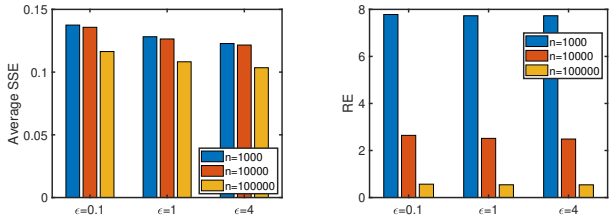


Fig. 5. Performance by varying population size

where the scheme generates a synopsis of the input dataset and the K-means clustering algorithm is executed on the synthetic data that is produced by the synopsis.

There are several works aimed at providing local privacy for K-means clustering analysis. The privacy concepts applied in the current literature mainly focus on local differential privacy. Nissim and Stemmer [25] proposed an algorithm for minimum enclosing ball, they make use of an LDP algorithm called GoodCenters. GoodCenters utilizes a locality sensitive hash function to hash the input points, which makes the collision happen to items with smaller distances with higher chance and further items with much less chance. Later Stemmer and Kaplan [26] propose an improved algorithm based on GoodCenters. The improved algorithm reduces the number of required interaction rounds significantly while at the same time reducing the multiplicative error. A follow-up work [27] further reduces the additive error. However, these works use a relaxed differential privacy definition, (ϵ, δ)-local differential privacy. Besides, Xia *et al.* [28] proposed a feature transformation method that encodes the user's data as a product of a binary string with a coefficient, which balances the accuracy, communication costs, and privacy very well. To enhance the privacy and hide the cluster each user belongs to in the intermediate rounds, the reports regarding the users' closest center is perturbed by the LDP protocol as well. A very recent work [29] provides an approximation algorithm for K-means clustering problem in one-round local differential privacy. They show that the proposed method achieves a similar small additive error when applied in the shuffled DP model, where a shuffler sits between the encoder and decoder. To our best knowledge, our work is the first to consider the d_χ-privacy in the K-means clustering problem. The d_χ-privacy related application has been discussed in Section II-C.

VI. Conclusion

In this paper, we considered the privacy problem of K-means clustering while providing local privacy to users' data. We adopted a generalized differential privacy definition, d_χ-privacy, which quantifies the distinguishability level based on the Euclidean distance between the data records. We proposed a bounded perturbation mechanism (BPM), which perturbs the data record as a whole which eliminates the need of splitting the privacy budget among different dimensions while outputting report in a bounded domain to maintain interpretability of the data and limit the bandwidth requirement. We formally proved that the proposed bounded perturbation mechanism achieves d_χ-privacy. Furthermore, we proposed an algorithm with an efficient expected complexity to sample from the proposed distribution of BPM and applied the BPM to K-means clustering. We theoretically analysed the proposed method and experimentally showed that the proposed method has a significant advantage compared to the traditional Laplace mechanism.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] W. Yang, K.-Y. Lam, J. Zhou, X. Luo, Q. Shen, and Z. Xu, "Automated cyber threat intelligence reports classification for early warning of cyber attacks in next generation soc." in *ICICS*, 2019, pp. 145–164.

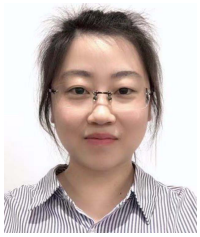
[2] R. Lakshmi and S. Baskar, "Efficient text document clustering with new similarity measures," *International Journal of Business Intelligence and Data Mining*, vol. 18, no. 1, pp. 49–72, 2021.

[3] S. Wang, X. Zhang, Y. Cheng, F. Jiang, W. Yu, and J. Peng, "A fast content-based spam filtering algorithm with fuzzy-svm and k-means," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2018, pp. 301–307.

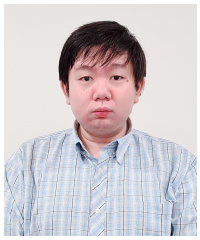
[4] R. Ghezelbash, A. Maghsoudi, and E. J. M. Carranza, "Optimization of geochemical anomaly detection using a novel genetic k-means clustering (gkmc) algorithm," *Computers & Geosciences*, vol. 134, p. 104335, 2020.

[5] M. G. Pradana and H. T. Ha, "Maximizing strategy improvement in mall customer segmentation using k-means clustering," *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 19–25, 2021.

[6] 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*, 2014, pp. 1054–1067.
[7] A. Differential Privacy Team, "Learning with privacy at scale," Available at: https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, accessed: May 19, 2023.
[8] Y. Zhao, J. Zhao, M. Yang, T. Wang, N. Wang, L. Lyu, D. Niyato, and K.-Y. Lam, "Local differential privacy-based federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8836–8853, 2020.
[9] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
[10] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2013, pp. 82–102.
[11] I. Wagner and D. Eckhoff, "Technical privacy metrics: a systematic survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–38, 2018.
[12] O. Feyisetan, T. Diethe, and T. Drake, "Leveraging hierarchical representations for preserving privacy and utility in text," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 210–219.
[13] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indis-tinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pp. 901–914.
[14] M. E. Muller, "A note on a method for generating points uniformly on ¡i¿n¡/i¿-dimensional spheres," *Commun. ACM*, vol. 2, no. 4, p. 19–20, Apr. 1959. [Online]. Available: https://doi.org/10.1145/377939.377946
[15] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the sulq framework," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2005, pp. 128–138.
[16] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private k-means clustering," in *Proceedings of the sixth ACM conference on data and application security and privacy*, 2016, pp. 26–37.
[17] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.
[18] T. Ni, M. Qiao, Z. Chen, S. Zhang, and H. Zhong, "Utility-efficient differentially private k-means clustering based on cluster merging," *Neurocomputing*, vol. 424, pp. 205–214, 2021.
[19] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "Gupt: privacy preserving data analysis made easy," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 349–360.
[20] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 2011, pp. 813–822.
[21] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett, "Privgene: differentially private model fitting using genetic algorithms," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 665–676.
[22] Z. Lu and H. Shen, "A convergent differentially private k-means clustering algorithm," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2019, pp. 612–624.
[23] ——, "Differentially private k k-means clustering with convergence guarantee," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 4, pp. 1541–1552, 2020.
[24] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, and H. Jin, "Differentially private k-means clustering and a hybrid approach to private optimization," *ACM Trans. Priv. Secur.*, vol. 20, no. 4, oct 2017. [Online]. Available: https://doi.org/10.1145/3133201
[25] K. Nissim and U. Stemmer, "Clustering algorithms for the centralized and local models," in *Algorithmic Learning Theory*. PMLR, 2018, pp. 619–653.
[26] U. Stemmer and H. Kaplan, "Differentially private k-means with constant multiplicative error," in *NeurIPS*, 2018.
[27] U. Stemmer, "Locally private k-means clustering," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2020, pp. 548–559.
[28] C. Xia, J. Hua, W. Tong, and S. Zhong, "Distributed k-means clustering guaranteeing local differential privacy," *Computers & Security*, vol. 90, p. 101699, 2020.
[29] A. Chang, B. Ghazi, R. Kumar, and P. Manurangsi, "Locally private k-means in one round," *arXiv preprint arXiv:2104.09734*, 2021.



Mengmeng Yang received her B.Eng and M.Sc degrees from Qingdao Agriculture University, China, in 2011 and Shenyang Normal University, China, in 2014 respectively, and a Ph.D. degree from Deakin University in Computer Science, Australia, in 2019. Dr Mengmeng Yang is currently a research fellow in Strategic Centre for Research in Privacy-Preserving Technologies and Systems (SCRiPTS), Nanyang Technological University.
Her research interests include privacy preserving, data mining, and network security.



Ivan Tjuawinata received his B.Sc. (First Class Honours) from Nanyang Technological University, Singapore, in 2012 and his Ph.D. degree from Nanyang Technological University, Singapore, in 2017.
Dr. Ivan Tjuawinata is currently a research fellow in Strategic Centre for Research in Privacy-Preserving Technologies and Systems (SCRiPTS), Nanyang Technological University.
His research interests include secure multiparty computation, secret-sharing scheme, and coding theory.

Kwok Yan Lam received his B.Sc. (First Class Honours) from the University of London in 1987 and his Ph.D. from the University of Cambridge in 1990. Professor Lam is the Associate Vice President (Strategy and Partnerships) and Professor in the School of Computer Science and Engineering at the Nanyang Technological University (NTU), Singapore. He is currently also the Executive Director of the Strategic Centre for Research in Privacy-Preserving Technologies and Systems (SCRiPTS), and Director of NTU's SPIRIT Smart Nation Research Centre. From August 2020, Professor Lam is also on part-time secondment to the INTERPOL as a Consultant at Cyber and New Technology Innovation. He served as the Director of the Nanyang Technopreneurship Center 2019-2022, and as Program Chair (Secure Community) of the Graduate College at NTU 2017-2019. Professor Lam has been a Professor of the Tsinghua University, PR China (2002-2010) and a faculty member of the National University of Singapore and the University of London since 1990.
His research interests include Distributed Systems, IoT Security Infrastructure and Cyber-Physical System Security, Distributed Protocols for Blockchain, Biometric Cryptography, Homeland Security, and Cybersecurity.