UC Davis

UC Davis Previously Published Works

Title

Differentially Private K-means Clustering Applied to Meter Data Analysis and Synthesis

Permalink

https://escholarship.org/uc/item/81p7r7nr

Journal

IEEE Transactions on Smart Grid, 13(6)

ISSN

1949-3053

Authors

Ravi, N

Scaglione, A

Kadam, S

et al.

Publication Date

2022

DOI

10.1109/TSG.2022.3184252

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Differentially Private K-means Clustering Applied to Meter Data Analysis and Synthesis

Nikhil Ravi, Anna Scaglione, Sachin Kadam, Reinhard Gentz,
Sean Peisert, Brent Lunghino, Emmanuel Levijarvi, and Aram Shumavon

arXiv:2112.03801v2 [eess.SP] 22 Apr 2022

***Abstract*—The proliferation of smart meters has resulted in a large amount of data being generated. It is increasingly apparent that methods are required for allowing a variety of stakeholders to leverage the data in a manner that preserves the privacy of the consumers. The sector is scrambling to define policies, such as the so called '15/15 rule', to respond to the need. However, the current policies fail to adequately guarantee privacy. In this paper, we address the problem of allowing third parties to apply K-means clustering, obtaining customer labels and centroids for a set of load time series by applying the framework of differential privacy. We leverage the method to design an algorithm that generates differentially private synthetic load data consistent with the labeled data. We test our algorithm's utility by answering summary statistics such as average daily load profiles for a 2-dimensional synthetic dataset and a real-world power load dataset.**

***Index Terms*—differential privacy, clustering, smart grids, summary statistics, synthetic load generation**

Nomenclature

X	The subset of the superset that is queried
P	The number of datapoints queried
$\boldsymbol{x}_p$	The p-th datapoint, e.g., power load of a house
d	The dimension of each data point
$\mathbb{q}(X)$	The query
$\boldsymbol{q}$	The query answer
$\tilde{\boldsymbol{q}}$	The Differentially private query answer
$\boldsymbol{c}_i$	The centroid of cluster i
ℓ_p	The label of point p
$r(X, \boldsymbol{q})$	The clustering loss
$\bar{r}$	The DP accuracy loss
ϵ	The privacy budget
ϵ_c	The privacy budget associated with the centroids
ϵ_ℓ	The privacy budget associated with the labels
δ	The privacy guarantee
δ_c	The privacy guarantee associated with the centroids
δ_ℓ	The privacy guarantee associated with the labels
$\Delta\mathbb{c}$	The sensitivity of the centroid query
$\Delta\mathbb{l}$	The sensitivity of the label query

I. Introduction

The growth in capabilities for data collection and computation has led to better products and greater market efficiencies in many sectors [1], [2]. In this paper, we consider the case of electric utilities that, over the last decade, have significantly expanded their residential metering and sensor deployments over distribution feeders responding to regulation for demand forecasting transactive energy, power flow optimization, fault-detection, planning for distributed energy resources (DERs), and for billing and automatic disconnection (see e.g., [3] for a taxonomy of the applications). An important commercial application for the data is to target consumers for promotional campaigns, mapping their behavior in classes to improve pricing or to provide incentives for reducing or shifting consumption, or for installing DER. Among the data queries that are useful to analyze customer data, *clustering* is one of the most common (see e.g., [4] for a survey) because of its unsupervised nature and relatively high accuracy. In [5], the authors present a variety of clustering techniques to identify typical *daily load profile* of consumers, and in [6], the authors propose a K-means clustering technique to identify similar types of load profiles for demand variation analysis and energy loss estimation.

None of the papers on clustering cited above addressed the issue of privacy in releasing the query results on customers' smart meters' database. Safeguarding against unintended disclosure of private data is critical in this sector. In fact, since the early deployments of residential smart meters, many researchers have investigated methods to maintain privacy while still enabling the data to be useful (see [7] for a survey). The most conventional methods in the industry are access control (e.g., see [8]), "anonymization" (e.g., see [9]), and data falsification techniques. Access control techniques alone allow all or no access. Anonymization masks data or makes it more general, but it has been shown repeatedly [10], that it is either removing too much of the data so that the query becomes useless or that it is typically insufficient, enabling linkage attacks that can be used to re-identify records. For example, researchers were able to re-identify some users in the anonymized data consisting of AOL search engine queries even when user IDs and IP addresses were removed [11], and similar approaches were used to re-identify anonymized records in the Netflix Prize dataset [12] and the Personal Genome Project [13]. A further disadvantage of anonymization is that once the data set is released, it is forever vulnerable to future reidentification attacks. For electric grid data, regulators

N. Ravi and A. Scaglione are with the Department of Electrical and Computer Engineering, Cornell Tech, e-mail: nr337@cornell.edu. S. Kadam is with Arizona State University. R. Gentz did this work while employed at Lawrence Berkeley National Laboratory. S. Peisert is with Lawrence Berkeley National Laboratory. B. Lunghino, E. Levijarvi, and A. Shumavon are with Kevala, Inc. The authors would also like to acknowledge Kaiya Levine.

This research was supported by the Director, Cybersecurity, Energy Security, and Emergency Response, Cybersecurity for Energy Delivery Systems program, of the U.S. Department of Energy, under contract DE-AC02-05CH11231. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors of this work.

have proposed several policies to share electric consumer data in the public domain. Of particular note is the "15/15 Rule" [14] which states that any aggregation of customer data is considered anonymous if it contains at least fifteen customers and if no single customer's data comprises 15% or more of the total values in the aggregated answer. There is no scientific rationale behind this rule; in fact, we show in Section II-A that it offers no privacy guarantee.

With this in mind, this paper presents an approach to applying differential privacy (DP) [15] for releasing the clustering results of the K-means algorithm to allow several stakeholders to query the meters' data while preserving privacy. We also use the mechanism as a stepping stone to publish differentially private synthetic data that emulate the consumer behavior in a class. DP consists of a growing suite of randomized methods to publish the output of data queries while guaranteeing that even multiple query answers are statistically unlikely to reveal information about an individual's data contributions, or lack thereof. Within the DP framework, there is no release of raw data, only the output of differentially private queries. Because DP acts as a "guard" between the query and response, DP is also capable of ceasing to respond to future queries once a pre-defined "privacy budget" has been reached. Another important aspect of DP methods is that they are tailored to the query and the database that is queried, which explains why our paper focuses on clustering. In contrast to anonymization techniques, DP mechanisms corrupt the query answers and thus, produce relatively lower accuracy, depending on the amount of noise added. However, the accuracy-privacy tradeoff is analytically quantifiable through DP.

Generic differentially private clustering techniques have been previously presented in [16]–[21]. The authors in [16] proposed a heuristic-based outlier-eliminated DP clustering mechanism with adaptive Laplacian noise. The authors in [17] proposed an iterative K-means clustering algorithm for data in high-dimensional Euclidean spaces. In [19], the authors proposed a local DP iterative clustering algorithm where noise is added at the user's end before transmitting the data to the aggregator. While guaranteeing DP, these techniques may not converge. Instead, the authors in [20] proposed a clustering algorithm that performs an input perturbation in each iteration, which offers convergence guarantees but drives the cost of DP higher depending on the number of iteration required for convergence. The authors of [18], [21] both propose methods with better initial point selection to improve the clustering accuracy, where they first generate $K' \gg K$ centroids by running the adaptive DP K-Means algorithm on randomly divided subsets of the dataset, before merging them into K centroids through an iterative process. These papers do not consider the privacy loss on the publication of cluster labels, which is ultimately what the classifier wants to know. Also, we note that the data such as a house's load profile are relatively smooth, and the i.i.d. noise added can be partly filtered out, whereas our proposed randomized mechanism leverages this fact to improve the performance.

Having classified the customers, synthetic load models that emulate the corresponding profiles are an important tool in power system studies to run realistic simulations. To overcome the shortage of publicly available large scale load datasets, researchers have tried to fill the gap by either publishing anonymized real data or creating synthetic datasets using historical datasets [22], [23]. Recently, researchers have adapted the use of conditional Generative Adversarial Networks (GAN) to generate realistic, synthetic week-long time-series load profiles at high resolutions [24], [25]. But these frameworks lack privacy guarantees for the historical data used in the process. To make matters worse, if the distribution of the generated synthetic data is close to the distribution of the true historic data, then any privacy leaked by the estimated statistics from the true data will also be leaked by the synthetically generated data. Also, we argue that the statistical structure of each class data can be approximated well by a multivariate log-normal distribution, whose generation is far less complex to train and use compared to a GAN.

A synopsis of our contributions is as follows:

1) We propose a novel differentially private clustering mechanism, which to the best of our knowledge contains:
   - The first analysis of the privacy leakage on the publication of the noisy labels and the first mechanism to publish them with DP guarantees.
   - The first of its kind, optimum scheme for adding *colored* Gaussian noise to the cluster centroids, that we recently proposed in [26]. This mechanism also provides greater accuracy for a given privacy budget compared to adding white noise, and also performs better in terms of privacy leakage compared to existing literature.
2) We also provide a mechanism that, leveraging the empirical good fit of Advanced Metering Infrastructure (AMI) data with a mixture of multivariate log-normal vectors, generates differentially private synthetic load data for each cluster, that can be safely published for simulation studies.

Finally, we test the efficacy of the proposed mechanism on samples drawn from a Gaussian mixture and a real-world AMI dataset. We wish to remark that the publication mechanism proposed for labels and centroids is broadly applicable to any type of data, while the generation of synthetic data relies on the AMI statistics and, thus, is domain specific.

Paper organization: In Section II, we introduce our problem statement, discuss the threat model, before introducing the DP framework. In Section III, we describe a DP mechanism for the publication of the clustering query and in Section IV, we present a model to generate synthetic load profiles. Finally, in Section V, we numerically test our algorithms, before concluding this paper in Section VI.

Notation: Boldfaced lower-case (upper-case respectively) letters denote vectors (matrices respectively) and x_i (X_{ij} respectively) denotes the i^{th} element of a vector $\boldsymbol{x}$ (the ij^{th} entry of a matrix $\boldsymbol{X}$ respectively). Calligraphic letters denote sets and $|\cdot|$ their cardinality. Finally, $[N]$ denotes the set of integers $\{0, 1, \ldots, N-1\}$.

II. Preliminaries

In the following, we denote by X a set of feature vectors $\boldsymbol{x}_p$, $p \in [P]$ embedded in $\mathbb{R}^d$ that are in a database $\mathcal{X}$, which we can organize as a $P \times d$ matrix $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P]^\intercal$.

Specifically, in an energy system that is considered in this paper, $\mathcal{X}$ is the set of homes that a utility company serves, $\boldsymbol{x}_p$ is the time-series of meter data of house p. To review the basic concepts, we set the problem in general terms and denote by $\mathbb{q}(X)$ the function that maps X onto the query answer, and denote the outcome by $\boldsymbol{q} \in \mathcal{Q}$, where $\mathcal{Q}$ is the domain of the query answers.

A. Is the "15/15" Rule sufficient?

For energy systems, the 15/15 rule focuses on an averaging query:

$$\mathbb{q}(X) = (1/P) \textstyle\sum_{x \in X} x, \tag{1}$$

where $P \geq 15$ and $\forall x \in X$, $x \leq 0.15\ \mathbb{q}(X)$. The claim of the 15/15 rule is that if one abides by it, then the data points are anonymized. However, there is a fundamental flaw with this type of aggregation when it comes to the privacy of each of the data records $x \in X$. Let us now consider an adversary who has queried for the average of the dataset X and a neighboring dataset X', which differ from X by just one point $x_0 \in X$, i.e., $X' = X \setminus \{x_0\}$. If both $\mathbb{q}(X)$ and $\mathbb{q}(X')$ are revealed to them, then by simple algebra:

$$x_0 = P\mathbb{q}(X) - (P-1)\mathbb{q}(X'). \tag{2}$$

Thus, they are able to infer the value of the point $x_0 \in X$ in the scenario described above, no matter what P is. This shows the need for a randomized response to the aggregate query, because responding with the exact answer to repeated queries will lead to privacy leakage. Next, we introduce our problem setup.

B. Problem Statement and Threat Model

In this paper, we consider the scenario where a trusted central data owner collects and stores the data and untrustworthy third parties query them. For example, an electric utility collects and stores the AMI data of the customers as seen in fig. 1. The electric utility agents, with legitimate access to the data, use them to perform operational functions such as scheduling, billing, etc., without being hindered by noisy data[1]. But the electric utility must abide by strict laws pertaining to customer privacy while publishing the data or aggregate query answers to external agents. Third party agents may query ($\mathbb{q}$) the utility's data, and receive a differentially privatized answer ($\tilde{\boldsymbol{q}}$) instead of the true query response ($\boldsymbol{q}$), as shown in fig. 2. The quality of the DP response depends on the budget $((\epsilon_c, \delta_c), (\epsilon_\ell, \delta_\ell))$ that the analyst is willing to spend to get the answer. The concept of privacy budget is explained in detail later in Section II-C.

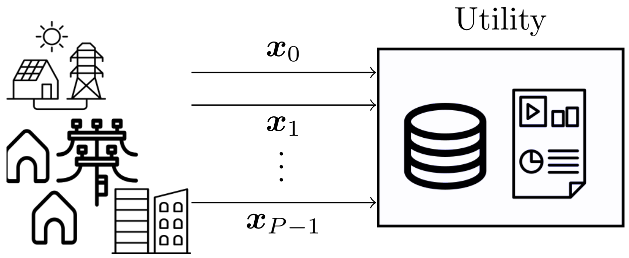

Fig. 1: The utility collects the raw AMI data from the customers it serves.

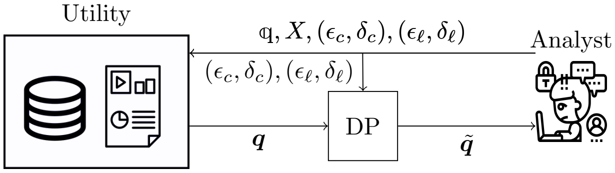
Utility
$\mathbb{q}, X, (\epsilon_c, \delta_c), (\epsilon_\ell, \delta_\ell)$
Analyst
$(\epsilon_c, \delta_c), (\epsilon_\ell, \delta_\ell)$
DP
$\boldsymbol{q}$
$\tilde{\boldsymbol{q}}$

Fig. 2: A third party analyst may query the dataset.

The goal of this paper is to study a differentially private randomized mechanism that would allow an analyst to perform a K-means clustering query on $X \subseteq \mathcal{X}$; that is, the analyst will obtain the information about the centroids as well as the labels of the data points in X through a randomized algorithm that will make it either difficult or impossible to tell if the database X or X' was queried, with X' missing one feature vector relative to X. The randomized mechanism proposed involves adding structured noise to the cluster centroids and to a subset of the labels. Prior to describing the mechanism, we review both the K-means clustering algorithm and the basic notion of DP next.

1) Review of K-means clustering: The K-means algorithm splits the dataset, into $K > 1$ subsets (clusters) and assigns a label to each point corresponding to the nearest cluster centroid to itself. In other words, the query we are interested in is given by $\mathbb{q} : \mathbb{R}^{P \times d} \to [K]^P \times \mathbb{R}^{K \times d}$. The problem can formally be posed as an optimization problem of the form:

$$\arg\min_{\mathfrak{C}} \ \frac{1}{P} \sum_{k \in [K]} \sum_{p \in \mathcal{C}_k} \|\boldsymbol{x}_p - \boldsymbol{c}_k\|_p \tag{3a}$$

$$\text{s.t.} \quad \bigcap_{k \in [K]} \mathcal{C}_k = [P], \tag{3b}$$

where $\mathfrak{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ is the partition in K clusters and $\boldsymbol{c}_k \in \mathbb{R}^d$ is the centroid of a cluster $\mathcal{C}_k$, obtained by averaging the points $\{\boldsymbol{x}_p\}_{p \in \mathcal{C}_k}$. The objective in eq. (3a) minimizes the cost of clustering assignment $\mathfrak{C}$ with the constraint in eq. (3b), so that every point in the database is assigned a cluster label whose centroid is the closest. The output of this algorithm consists of two components:

$$\mathbb{q}(X) := \{\mathbb{c}(X), \mathbb{l}(X)\}, \tag{4}$$

where

1) $\mathbb{c}(X)$ – with outcome $\boldsymbol{c} = [\boldsymbol{c}_0^\intercal, \ldots, \boldsymbol{c}_{K-1}^\intercal]^\intercal$ is the sub-query that returns the set of K centroids $\{\boldsymbol{c}_k \in \mathbb{R}^d\}_{k \in [K]}$.
2) $\mathbb{l}(X)$ – with outcome $\boldsymbol{\ell}$ is the sub-query that assigns a label to each feature point $\boldsymbol{x}_p$ that corresponds to an index of such centroids.

The performance of the K-means query outcome, $\boldsymbol{q} = \{\boldsymbol{c}, \boldsymbol{\ell}\}$, is measured in terms of the ***clustering loss*** which is defined as:

[1] There exists a second school of DP works, including [27], called Local DP, where there is no central data owner with access to the raw data. Here, individual customers add local noise to their own data before forwarding it to the data owner. In such scenarios, there is a significant cost to the operational functions of the electric utility, as it now does not possess the raw data it needs.

Definition 1 (Clustering Loss)**.** *The clustering loss of a K-means clustering algorithm is given by the sum of squares of the distance from each point to the centroids of its cluster, i.e.,*

$$r(X, \boldsymbol{q}) := \frac{1}{P} \sum_{p \in [P]} \|\boldsymbol{x}_p - \boldsymbol{c}_{\ell_p}\|_2^2, \quad \textit{where} \quad \boldsymbol{q} = \{\boldsymbol{c}, \boldsymbol{l}\} \quad (5)$$

We note that this is equivalent to the minimizer of the clustering query in eq. (3a), and that a lower clustering loss is better. In later sections, we introduce the DP mechanism that will corrupt the query outcome via an additive noise mechanism, and publish $\tilde{\boldsymbol{q}}$ in place of $\boldsymbol{q}$. This naturally induces an additional loss on top of the clustering loss. Thus, we define the ***DP (query) accuracy loss*** as follows:

Definition 2 (DP (query) Accuracy Loss)**.** *Given a query answer $\boldsymbol{q}$ and a DP answer $\tilde{\boldsymbol{q}}$, the DP query accuracy loss, or simply the DP accuracy loss, is given by:*

$$\bar{r}(X, \boldsymbol{q}, \tilde{\boldsymbol{q}}) = \frac{r(X, \tilde{\boldsymbol{q}}) - r(X, \boldsymbol{q})}{r(X, \boldsymbol{q})}, \quad (6)$$

where $r(X, \tilde{\boldsymbol{q}})$ is the clustering loss if the outcome of the K-means clustering algorithms were $\tilde{\boldsymbol{q}}$. Since $\boldsymbol{q}$ is the minimizer of eq. (3), $\bar{r} \geq 0$, and the equality is met in the absence of a DP mechanism.

It is also important to define the *reward* of our DP query answer. The addition of noise strengthens the privacy guarantees while increasing the DP accuracy loss. Since the reward and DP accuracy loss have a negative correlation, we will discuss the performance of our mechanisms in terms of their DP accuracy loss alone, with the understanding that lower the DP accuracy loss, the higher is the reward.

Finally, to restate, the goal of our mechanism is to guarantee differential privacy of the K-means clustering query answer while minimizing the DP accuracy loss.

2) Threat model to the K-Means Clustering Query: As discussed in the prior sections, traditional rules of thumb adapted by specific industries have flawed quantification of privacy guarantees, and anonymization often fails in the presence of substantial side information. In this regard, the threat we consider in this paper is that of a third-party analyst's ability to discern the value of any particular data point in the dataset that is under investigation.

To illustrate the threat, consider a scenario when both the centroids and labels are readily available for a dataset X. Now, suppose an additional point x is added to the dataset and this leads to a change in a single centroid, say of a cluster k. Notice that each centroid is similar to the average query that was under investigation in Section II-A. Now, using the information about the population of each cluster obtained from the vector of labels, an adversary can infer the value of the point x as described in Section II-A.

This threat is especially relevant in smart grid data analysis. To elaborate, consider the average query and a dataset that contains power loads of $P-1$ houses without solar photovoltaics (PVs) and one house with solar PVs installed. During the daytime, there will be periods during which the latter does not consume any power from the electric grid, and in fact injects power to the grid. Then, its load measurement will be negative (by convention) as opposed to the positive load measurements of all the other houses in the dataset. Thus, the average query with and without the last house will be vastly different from one another, making that house highly sensitive to the average query and a possible source of privacy leakage, as now the fact that it contains a solar PV can be inferred by the analyst.

Remark 1 (Internal and External Threats)**.** *We do not consider insiders (of the organization that stores the data) with legitimately acquired access to the data as threats. Instead, we are concerned with the inference of a data point's involvement after a particular aggregate query has been published to an external untrustworthy third party.*

C. Differential Privacy

We now introduce DP as an alternative to the widely used "15/15" Rule in the smart grid industry to publish private aggregate data. We denote the DP query answer by $\tilde{\mathbb{q}}(X)$, and has a random outcome $\tilde{\boldsymbol{q}} \in \mathcal{Q}$, with distribution $f(\tilde{\boldsymbol{q}}|X)$ (the probability density function for continuous random queries and the probability mass function for discrete random variables). We briefly introduce the conventional definitions that explain how differential privacy is measured and established. The first and the most widespread definition of differential privacy was introduced in [15], [28]. It states that:

Definition 3 ((ϵ, δ)-Differential privacy)**.** *A randomized mechanism $\tilde{\mathbb{q}}$ is (ϵ, δ)-differentially private if for all neighboring datasets X and X' that differ in one point, for any arbitrary event pertaining to the outcome of the query, the randomized mechanism satisfies the following inequality*

$$\forall \mathcal{S}, \quad Pr(\tilde{\mathbb{q}}(X) \in \mathcal{S}) \leq \exp(\epsilon) Pr(\tilde{\mathbb{q}}(X') \in \mathcal{S}) + \delta, \quad (7)$$

where $Pr(\mathcal{A})$ denotes the probability of the event $\mathcal{A}$, for some privacy budget $\epsilon \geq 0$ and $\delta \in [0, 1]$.

Note that, since δ is a bound that may not be tight, smaller values of δ are possible. Hence, (ϵ, δ) guarantees are sufficient but not necessary conditions to ensure that information about X leaks. The authors in [29] introduced a revised definition of privacy as follows:

Definition 4 ((ϵ, δ)-Probabilistic Differential privacy)**.** *The so-called privacy leakage function $L_{XX'}$ is the log-likelihood ratio between the two hypotheses that the query outcome $\tilde{\boldsymbol{q}}$ is the answer generated by the data X or the data X' that differ by one element. Mathematically:*

$$L_{XX'}(\tilde{\boldsymbol{q}}) := \log \frac{f(\tilde{\boldsymbol{q}}|X)}{f(\tilde{\boldsymbol{q}}|X')}, \quad (8)$$

A randomized mechanism $\tilde{\mathbb{q}}(X)$ is (ϵ, δ) differentially private for X if and only if:

$$\sup_{X'} \; Pr\left(L_{XX'}(\tilde{\boldsymbol{q}}) > \epsilon\right) \leq \delta. \quad (9)$$

It can be shown that (ϵ, δ)-PDP is a strictly stronger condition than (ϵ, δ)-DP.

Theorem 1 (PDP implies DP [30])**.** *If a randomized mechanism is (ϵ, δ)-PDP, then it is also (ϵ, δ)-DP, i.e.,*

$$(\epsilon, \delta) - PDP \Rightarrow (\epsilon, \delta) - DP, \text{ but } (\epsilon, \delta) - DP \not\Rightarrow (\epsilon, \delta) - PDP.$$

In dealing with a multidimensional answer $\tilde{\boldsymbol{q}}$ (as is for the case of a k-means clustering algorithm that returns K centroids and P data labels) a first common simplification is to use independent mechanisms for different components; a second common simplification is to use the following lemma to map the (ϵ_j, δ_j) results for the scalar independent mechanism employed onto a global (ϵ, δ) result, rather than $L_{xx'}(\tilde{\boldsymbol{q}}) = \sum_{j=1}^{m} L_{xx'}(\tilde{q}_j)$.

Lemma 1 (Sequential composition [31])**.** *If n randomized algorithms M_i, $i = 1, 2, \ldots, n$, are (ϵ_i, δ_i)-DP, then the sequential execution of these algorithms on the database X provides $(\sum_i \epsilon_i, \sum_i \delta_i)$-differential privacy.*

In this paper, we use Definition 4 which has an immediate statistical interpretation: if values of (ϵ, δ) are close to zero, even when one adopts the optimum statistical test for the hypotheses that the randomized answer is produced by the neighboring datasets X or X' that differ in one point, the test produces results that mostly are incorrect or are unreliable answers (i.e., there is a non-zero probability that the test is wrong). Of course, this comes at a cost in terms of accuracy of the answer, which is important to account for.

Next, we describe the mechanism $\tilde{\mathbb{q}}(X) = \{\tilde{\mathbb{c}}(X), \tilde{\mathbb{l}}(X)\}$ that renders the clustering query answer (ϵ, δ)-DP.

III. An (ϵ, δ)-DP Mechanism for K-Means Clustering

The K-means clustering query, as described in II-B1, has two components: the centroids and the labels $\mathbb{q}(X) = \{\mathbb{c}(X), \mathbb{l}(X)\}$. Our method introduces independent randomized mechanisms applied to each component and uses composition as defined in Lemma 1 to provide overall DP guarantees.

A. Differential Privacy Guarantees for Cluster Centroids

In this section, we first remind the readers of the classical white Gaussian noise mechanism and then introduce the novel additive colored Gaussian noise mechanism, where we perturb the output by adding noise to the cluster centroids found by an appropriate clustering algorithm with convergence guarantees.

As discussed in Section II-A, any DP mechanism should introduce uncertainty on the query answer so that its random outcomes under two neighboring datasets X or X' are statistically indistinguishable with high probability.

1) Review of the White Gaussian Noise Mechanism: The Gaussian noise output perturbation mechanism is a popular option in DP for publishing a variety of statistics. The approach entails adding i.i.d. noise to the cluster centroids as follows:

$$\tilde{\mathbb{c}}(X) = \mathbb{c}(X) + \boldsymbol{\eta} = \boldsymbol{c} + \boldsymbol{\eta}, \tag{10}$$

where $\boldsymbol{\eta}$ is a normally distributed noise vector with mean $\mathbf{0}$ and variance $\sigma^2 \boldsymbol{I}$. The following theorem provides the DP guarantees for such a method:

Theorem 2 (Cluster centroids are (ϵ_c, δ_c)-DP [15])**.** *The additive noise mechanism in eq. (10) provides (ϵ_c, δ_c)-DP for any two neighboring datasets X and X' when $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, for some $\epsilon_c \geq 0$ and $\delta_c \in [0, 1]$, where*

$$\sigma \geq \frac{\Delta \mathbb{c}}{\epsilon_c} \sqrt{2 \log(2/\delta_c)} \tag{11}$$

and $\Delta \mathbb{c}$ is the query sensitivity given by:

$$\Delta \mathbb{c} = \sup_{X'} \|\mathbb{c}(X) - \mathbb{c}(X')\|_2.$$

But it can be shown that a white noise mechanism applied to a smooth time-series dataset can be filtered out. For example, in fig. 3, we plot the average of the daily power loads of the houses in a feeder along with a DP white Gaussian noise treated load curve, where noise standard deviation was set at $\sigma = 0.75$. This noise can easily be filtered out by a Saviztky-Golay filter [32] (of order 1 with a window size of 300), a generalized moving average mechanism, as shown in bottom plots in fig. 3. For a dataset containing the power loads of houses in a set of feeders, the centroids found through a K-means algorithm are often in the low-frequency domain, and adding white noise that reside in the high-frequency domain is often not enough.

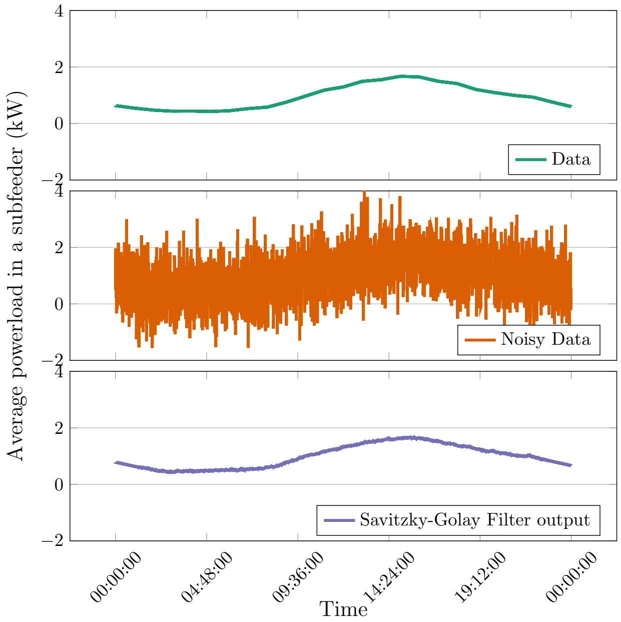

Fig. 3: Illustration of the pitfalls of white noise mechanisms on smooth time-series data.

2) Proposed Colored Gaussian Noise Mechanism: As an alternative to the conventional method of adding i.i.d. noise to each entry of the query, we consider the fact that, particularly in the case of time series, it is often the case that they are sparse in a transform domain, such as the Discrete Fourier Transform or the Wavelet transform. In such a scenario, we propose the addition of colored Gaussian noise instead, where the noise added depends on the relative positions of the centroids in question:

$$\tilde{\mathbb{c}}(X) = \mathbb{c}(X) + \hat{\boldsymbol{\eta}} \quad \text{where} \quad \hat{\boldsymbol{\eta}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}^{-1}), \tag{12}$$

where where $\boldsymbol{\Gamma}$ is the noise precision matrix. The following theorem states the privacy guarantees of this mechanism, with its proof presented in [26, Theorem 4].

Theorem 3 (Colored Gaussian Noise Mechanism is (ϵ_c, δ_c)-DP)**.** *The additive noise mechanism in eq. (12) provides (ϵ_c, δ_c)-DP for any two neighboring datasets X and X' that differ in one point, for some $\epsilon_c \geq 0$ and $\delta_c \in [0, 1]$.*

While the proof is presented in [26], we briefly provide the sketch of the proof here.

Proof. The design of the optimal noise vector hinges on the design of its covariance matrix. This is seen by analyzing the numerator of the DP accuracy loss function in eq. (6), which quantifies the deviation of the accuracy loss function (defined in eq. (5)) at the noisy query answer $\tilde{\boldsymbol{q}}$ from its optimal value obtained at the true query answer $\boldsymbol{q}$. Under the Gaussian mechanism, the clustering loss is given by:

$$\begin{aligned} r(X, \tilde{\boldsymbol{q}}) &= \frac{1}{P}\sum_{p\in[P]} \|\boldsymbol{x}_p - \tilde{\boldsymbol{c}}_{\ell_p}\|_2^2 = \frac{1}{P}\sum_{p\in[P]} \|\hat{\boldsymbol{\eta}}_{\ell_p} - \left(\boldsymbol{x}_p - \boldsymbol{c}_{\ell_p}\right)\|_2^2 \\ &= r(X, \boldsymbol{q}) + \sum_{k\in[K]} \|\hat{\boldsymbol{\eta}}_k\|_2^2 - \frac{2}{P}\sum_{p\in[P]} \langle \hat{\boldsymbol{\eta}}_{\ell_p}, \boldsymbol{x}_p - \boldsymbol{c}_{\ell_p}\rangle \\ &= r(X, \boldsymbol{q}) + \|\hat{\boldsymbol{\eta}}\|_2^2 - \frac{2}{P}\sum_{p\in[P]} \langle \hat{\boldsymbol{\eta}}_{\ell_p}, \boldsymbol{x}_p - \boldsymbol{c}_{\ell_p}\rangle, \end{aligned}$$

where $\hat{\boldsymbol{\eta}} = [\hat{\boldsymbol{\eta}}_0^T, \ldots, \hat{\boldsymbol{\eta}}_{K-1}^T]^T$ with $\hat{\eta}_k$ being the noise vector added to the centroid of cluster k. Thus, we can write the following:

$$\mathbb{E}_{\hat{\boldsymbol{\eta}}}[r(X, \tilde{\boldsymbol{q}}) - r(X, \boldsymbol{q})] = \mathrm{Tr}(\mathrm{covar}(\hat{\boldsymbol{\eta}})) = \mathrm{Tr}(\boldsymbol{\Gamma}^{-1}),$$

where $\mathrm{Tr}(\boldsymbol{A})$ is the trace of the matrix $\boldsymbol{A}$ given by the sum of the elements on its diagonal. Hence, in order to achieve the lowest DP accuracy loss, we have to minimize the trace (Tr) of the noise covariance matrix (the inverse of $\boldsymbol{\Gamma}$).

Concurrently, in order to maintain the DP guarantees, we need to satisfy eq. (11). Thus, we have the following condition:

$$\Delta\mathbb{c}^2 \leq \frac{\epsilon_c^2}{2\log(2/\delta_c)} =: \gamma_c, \tag{13}$$

where σ is set to 1, and the cluster query sensitivity is:

$$\Delta\mathbb{c} = \sup_{X'} \|\boldsymbol{\Gamma}^{1/2}\left(\mathbb{c}(X) - \mathbb{c}(X')\right)\|_2. \tag{14}$$

Thus, for all neighboring datasets $X' \in \mathcal{X}$ of X that differ in one element, we have that:

$$\left(\mathbb{c}(X) - \mathbb{c}(X')\right)^\intercal \boldsymbol{\Gamma}\left(\mathbb{c}(X) - \mathbb{c}(X')\right) \leq \Delta\mathbb{c}^2 \leq \gamma_c.$$

Finally, we write the following optimization problem to find the optimal noise covariance matrix that finds the covariance matrix that satisfies the DP conditions while providing the least DP accuracy loss:

$$\begin{aligned} &\min_{\boldsymbol{\Gamma}} \ \mathrm{tr}(\boldsymbol{\Gamma}^{-1}) \\ &\ \text{s.t.} \ \left(\mathbb{c}(X) - \mathbb{c}(X')\right)^\intercal \boldsymbol{\Gamma}\left(\mathbb{c}(X) - \mathbb{c}(X')\right) \leq \gamma_c, \forall X' \in \mathcal{X}. \end{aligned} \tag{15}$$

In the following lemma, we provide the means to find the optimal noise covariance (see [26, Lemma 5] for proof):

Lemma 2 (Optimal Choice of $\boldsymbol{\Gamma}$)**.** *Let the matrix $\boldsymbol{C}_{XX'}$ contain as its columns all possible $(\mathbb{c}(X) - \mathbb{c}(X'))$, $X' \in \mathcal{X}$ and let us assume that $\boldsymbol{C}_{XX'}$ is full row rank, and that the first Kd columns of $\boldsymbol{C}_{XX'}$, corresponding to the set $\mathcal{D} \subseteq \mathcal{X}$ have the smallest norms and are linearly independent, forming the matrix we refer to as $\boldsymbol{C}^*_{XX'}$. Then, the optimization problem in eq. (15) has a unique solution and it evaluates to:*

$$\boldsymbol{\Gamma}^\star = \boldsymbol{R}_{\boldsymbol{\lambda}^\star}^{-\frac{1}{2}}, \tag{16}$$

where $\boldsymbol{\lambda}^*$ is the vector consisting of the Lagrangian multipliers associated with the optimization problem and has only Kd non-zero values which correspond to the constraints associated with the set $\mathcal{D}$, and:

$$\boldsymbol{R}_{\boldsymbol{\lambda}^\star} := \sum_{X'\in\mathcal{D}} \lambda^\star_{X'}\left(\mathbb{c}(X) - \mathbb{c}(X')\right)\left(\mathbb{c}(X) - \mathbb{c}(X')\right)^\intercal, \tag{17}$$

where $\lambda^\star_{X'}, \forall X' \in \mathcal{D}$ are the non-zero Lagrange multipliers for the problem in eq. (15). Their values are:

$$\lambda_i^* = v_i^{-2} \ \ \textit{where} \ \ \boldsymbol{v} = \gamma_c \boldsymbol{M}^{-1}\boldsymbol{1}, \ \ M_{ij} = [\boldsymbol{C}^{*\frac{\intercal}{2}}_{XX'}]^2_{ij}. \tag{18}$$

Thus, the mechanism in eq. (12) is (ϵ_c, δ_c)-DP with the noise covariance set according to eq. (16). □

B. Differential Privacy Guarantees for the Labels

In this section, we focus on a mechanism for publishing the labels in a way that is differentially private. This task is less straightforward than the previous one. To begin with, if the number of points P is extremely high compared to the number of clusters K, then the removal of one point from the dataset might not necessarily change the labels of the rest of the points, it is unnecessary to randomize the labels. We have to ensure that we randomize the labels of a subset of the points, which when removed changes the labels of the rest of the points, i.e., those point which are sensitive to the label query. Secondly, to complicate matters, adding random errors to the labels of points very close to the centroids would result in extremely inaccurate results, making the answers completely useless. Given these two caveats, the task of any algorithm is two-fold:

i) Choose the ideal subset of points whose labels are to be randomized, and

ii) randomize the labels of these points such that we achieve the least error.

1) Choice of Points to Randomize: In a densely populated field of points, the effect of the removal of one point from the dataset on the position of the centroids is minimal. Thus, points closer to their centroids and points in clusters far away from other clusters tend to retain their labels. However, the labels of the points in the periphery of a cluster, and especially those in the vicinity of points from another cluster, might flip on the removal of a point from the dataset. Thus, only these edge points contribute to the sensitivity of the label query. The mechanism that we propose in this section only adds label noise to a subset of the population, say $\mathcal{L} \subseteq [P]$, which is the set of all the points whose label changes if any other point is removed. The composition of $\mathcal{L}$ is illustrated in fig. 4, where we have a set of points $\{0, \ldots, 19\}$. Here, when

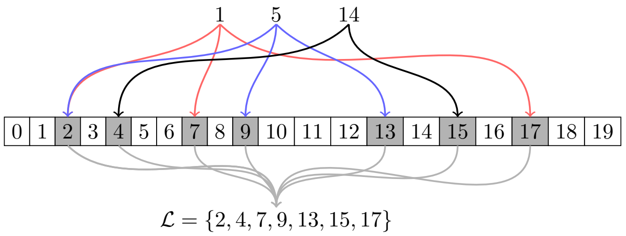

Fig. 4: The composition of the set of points whose labels are treated with the DP Label mechanism.

point 1 is removed from the set, the new cluster affiliations of the points in $\{2, 7, 17\}$ are different from what they were before. Similarly, the cluster affiliations of points in $\{2, 9, 13\}$ (respectively $\{4, 15\}$) are changed when point 5 (respectively point 14) is removed. The union of all such sets of points forms the set $\mathcal{L}$.

2) Randomized Mechanism for the Labels: Unlike in the case of cluster centroids, the use of Gaussian (or any other unbounded noise) mechanism is impermissible for the labels, as they are integers belonging to the set $[K]$. Thus, we employ a modulo-K additive mechanism where the output is given by:

$$[\tilde{\mathbb{l}}(X)]_p = \begin{cases} [\mathbb{l}(X)]_p \oplus_K \nu_p, & p \in \mathcal{L} \\ [\mathbb{l}(X)]_p, & \text{otherwise}, \end{cases} \tag{19}$$

where $\{\nu_p\}_{p\in\mathcal{L}}$ are i.i.d random variables in the set $[K]$ and $\oplus_K$ is the modulo K addition operator, which ensures that the range remains in the permissible set $[K]$. The mechanism involves the modular addition of i.i.d noise samples to the labels of each of the points in $\mathcal{L}$. In the following theorem, we provide the DP guarantees afforded by this mechanism with the proof in Appendix A.

Theorem 4 (Labels are $(\epsilon_\ell, \delta_\ell)$-DP)**.** *The modulo additive noise mechanism in eq.* (19) *provides* $(\epsilon_\ell, \delta_\ell)$*-DP for any two neighboring datasets* X *and* X' *when* $\{\nu_p\}_{p\in\mathcal{L}}$ *are i.i.d and have the following probability mass function:*

$$f(\nu_p) = \begin{cases} 1 - \rho, & \nu_p = 0 \\ \frac{\rho}{K-1}, & \nu_p = [K] \setminus \{0\}, \end{cases}$$

where $K \geq 2$, $\rho < 0.5$, $\Delta\mathbb{l}$ *is the query sensitivity given by:*

$$\Delta\mathbb{l} = \sup_{X'} \|\mathbb{l}_{\mathcal{L}}(X') - \mathbb{l}_{\mathcal{L}}(X)\|_0,$$

and δ_ℓ *is* 0 *when* $\epsilon_\ell > \Delta\mathbb{l} \times \log\left(\frac{(1-\rho)(K-1)}{\rho}\right)$ *and when* $\epsilon_\ell < \Delta\mathbb{l} \times \log\left(\frac{(1-\rho)(K-1)}{\rho}\right)$*, it is given by*

$$\delta_\ell = \Delta\mathbb{l}! \left[\frac{\rho(K-2)}{K-1}\right]^{\Delta\mathbb{l}} \times \sum_{m_c=0}^{\Delta\mathbb{l}} \sum_{m_0=\ell+m_c}^{\Delta\mathbb{l}} \frac{\left[\frac{(K-1)(1-\rho)}{\rho}\right]^{m_0} (K-2)^{-(m_0+m_c)}}{m_0! m_c! (\Delta\mathbb{l} - m_0 - m_c)!}.$$

We measure the label mechanism's degradation of the solution as the number of points whose labels are changed, i.e., $\|\mathbb{l}_{\mathcal{L}}(X) - \tilde{\mathbb{l}}_{\mathcal{L}}(X)\|_0$, and in expectation this loss is given by:

$$\mathbb{E}[\|\mathbb{l}_{\mathcal{L}}(X) - \tilde{\mathbb{l}}_{\mathcal{L}}(X)\|_0] = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{p\in\mathcal{L}} \mathbb{1}_{\{\nu_p \neq 0\}}\right] = |\mathcal{L}|\rho.$$

Finally, the total degradation of the combined centroid and label perturbations is measured as follows:

$$r(X, \tilde{\boldsymbol{q}}) := \frac{1}{P} \sum_{p\in[P]} \|\boldsymbol{x}_p - \tilde{\boldsymbol{c}}_{\tilde{\ell}_p}\|_2^2.$$

C. Mechanism and its Differential Privacy Guarantees

In Sections III-A to III-B, we provided differential private mechanisms for the publication of cluster centroids and point labels. In algorithm 1, we provide the overall DP K-means mechanism with colored Gaussian noise for the centroids and discrete noise for the labels. In the following corollary, we state the overall DP guarantees that encapsulates the complete query as described in eq. (4).

Algorithm 1 DP Mechanism with colored Gaussian noise and randomized labels

```
procedure DPK-MEANS($\mathbb{q}, X, \epsilon_c, \delta_c, \epsilon_\ell, \delta_\ell$)
    $\boldsymbol{q}(\boldsymbol{c}, \boldsymbol{\ell}) \leftarrow$ Cluster $X$ into $K$ clusters.
    Calculate $\boldsymbol{\Gamma}$ according to eq. (16) with $\epsilon_c$ and $\delta_c$.
    $\tilde{\boldsymbol{c}} \leftarrow \boldsymbol{c} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$.
    $\tilde{\boldsymbol{\ell}} \leftarrow \boldsymbol{\ell} \oplus \boldsymbol{\nu}$, with $\boldsymbol{\nu}$ distributed according to the conditions of theorem 4.
end procedure
 Output: $\tilde{\boldsymbol{q}} = \{\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\ell}}\}$.
```

Corollary 1 (Clustering query is (ϵ, δ)-DP)**.** *For any two neighboring datasets* X *and* X'*, the mechanisms described in eq.* (10) *and eq.* (19) *together provide* (ϵ, δ)*-DP guarantees with:*

$$\epsilon = \epsilon_c + \epsilon_\ell, \quad \delta = \delta_c + \delta_\ell. \tag{20}$$

The proof follows from Theorems 3 to 4 and Lemma 1.

IV. Use Case: Power Systems Aggregate Query Publication and Synthetic Load Generation

The guarantees provided by the proposed DP clustering mechanism are tailor-made to the ones that regulatory agencies are trying to achieve with regard to the publication of consumer data, since third-party agencies will not be able to gleam information about the presence, or lack thereof, of any individual entry in the database.

The method proposed is general and can be applied to any kind of data. In this section, we go one step further and use the guarantees afforded in the publication of labels and centroids to generate synthetic load profiles for the nodes in the dataset, leveraging an empirical property that is specific to AMI data. In fact, it so happens that the daily load profiles when divided into clusters (to be discussed in section V-B in detail) exhibit an excellent fit with a multivariate log-normal distribution whose generation is relatively simple. The first step to generate synthetic samples is to group the dataset by clusters, take the logarithm (adding a large constant if the net load data have negative values) and then use the DP centroids and DP covariance to fit separate Gaussian distributions to each cluster and then go back to power loads profiles taking an exponential (and subtracting the constant, if needed. Letting $\exp(X)$ (resp. $\log(\boldsymbol{X})$) denote the matrix where each element is the exponent (resp. logarithm) of the corresponding element in $\boldsymbol{X}$, the algorithm in algorithm 2 generates prototypical load shapes for each cluster. Our DP clustering mechanism renders the estimates of the means and labels differentially private. To obtain a DP covariance, we apply the Wishart mechanism [33]. Using this fitted distribution, we then generate multivariate Gaussian samples, before finally outputting the exponential of the generated samples. This approach is summarized in algorithm 2, where the α parameter is used to ensure that the argument of the log operation remains positive, and $\boldsymbol{\eta}_k$ and $\boldsymbol{\Phi}_k$ are the noise added to the cluster centroid and to the covariance matrix of the data pertaining to the cluster k.

Algorithm 2 Algorithm to generate Load Profiles

```
procedure LOGNORMSAMPLES
    Cluster $X$ into $K$ clusters.
    for $k \in [K]$ do
        $\boldsymbol{X}_k \leftarrow \log([\boldsymbol{x}_{p_1} \dots \boldsymbol{x}_{p_{|\tilde{\mathcal{C}}_k|}}] + \alpha \mathbf{1}\mathbf{1}^\intercal)$
        $\hat{\boldsymbol{\mu}}_k = \frac{1}{|\tilde{\mathcal{C}}_k|} \boldsymbol{X}_k \mathbf{1} + \boldsymbol{\eta}_k$
        $\hat{\boldsymbol{\Sigma}}_k = \frac{1}{|\tilde{\mathcal{C}}_k|} \boldsymbol{X}_k \boldsymbol{X}_k^\intercal + \boldsymbol{\Phi}_k$
        $\hat{\boldsymbol{X}} \leftarrow$ samples drawn from $\mathcal{N}(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$
        $\hat{\boldsymbol{X}} \leftarrow \exp(\hat{\boldsymbol{X}}) - \alpha \mathbf{1}\mathbf{1}^\intercal$.
    end for
end procedure
```

V. Numerical Results

In this section, the proposed methods are tested on a synthetic dataset consisting of points drawn from a Gaussian mixture and then with two real-world datasets.

A. Synthetic Dataset

In this section X consists of $P = 1000$ points, each randomly drawn from a Gaussian mixture with $K = 6$ equally likely components that are in $\mathbb{R}^2$ and are $\mathcal{N}(\boldsymbol{\mu}_k, \sigma \boldsymbol{I})$ $k \in [K]$ with $\boldsymbol{\mu}_k \in \mathbb{R}^2$. Note that because the mixture components covariance is $\boldsymbol{I}$, here we do not need colored noise for the centroids. The output of the K-means clustering algorithm is shown in fig. 5 and it is the clustering assignment that we consider as the true assignment. In the same figure, we highlight the elements of the set $\mathcal{L}$ by marking them with markers encircled by yellow borders. These points have additive noise added to their labels, as shown in eq. (19), and those whose labels changed after the DP mechanism are highlighted with square (□) markers with blue borders. In fig. 6, we show the variation of noise variances for various δ_cs, the variation of δ_c for various noise variances and the DP accuracy loss for various δ_cs. As expected, in order to achieve a stronger privacy guarantee (i.e., lower δ_c values for given σ), we require a higher privacy budget and, in a similar vein, a lower DP accuracy loss (in other words, a lower noise variance) requires a higher privacy budget for a given level of privacy guarantee. It is also important to note the effect of the sensitivity on the performance of the DP mechanism proposed for the labels. Queries that are highly sensitive to the underlying dataset require a larger noise variance and, in turn, reduce the reward of the DP query answer. This is illustrated in fig. 6 (right), where, as the sensitivity $\Delta \mathbb{l}$ increases, the expected percent of errors, $\rho|\mathcal{L}| * 100/P$, also increases for a given privacy budget.

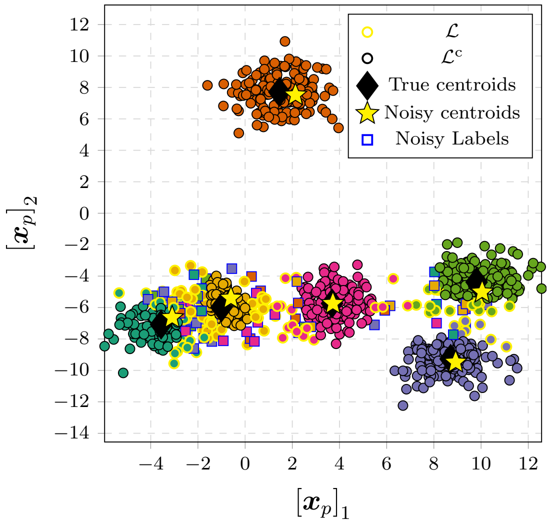

Fig. 5: Gaussian Mixture Dataset scatter plots: the six classes are indicated using different colors, with the points belonging to the set $\mathcal{L}$ indicated by smaller marks with a yellow border. Of these, the points whose labels changed are indicated using square (□) markers.

B. Real-World Power Systems AMI Dataset

In this section, we apply our mechanisms for the publication of K means statistics for $P = 1416$ houses' daily profiles (with one hour-resolution) that belong to 12 distribution circuits across California, USA. The dimension of each house's daily profile, $\boldsymbol{x}_p$, is $d = 25$ accounting for the load consumed from midnight to midnight. With a choice of $K = 6$ clusters, to visualize the daily profiles that are in $\mathbb{R}^{25}$, we map them on $\mathbb{R}^2$ using Multidimensional Scaling[2] (MDS) and, as before, in fig. 7, the points in $\mathcal{L}$ are highlighted using markers encircled in yellow borders. As seen in the MDS plot in fig. 7, we have close to 7 points that are clear outliers in the dataset. These outliers drive up the sensitivity value, which would require either higher noise covariance or a very large DP budget.

Domain Specific Knowledge: It is extremely important to pair the guarantees provided by DP with domain specific knowledge in order to extract the best possible reward. With the assumption that the data holder has knowledge regarding the class (i.e., commercial, residential, farm, etc.) to which

[2] MDS is a form of non-linear dimensionality reduction which is used to translate information about the pairwise 'distances' among a set of n objects or individuals into a configuration of n points mapped into an abstract Cartesian space [34]. It is important to note that points shown in MDS are solely for visualization purposes and not to be interpreted as containing behind-the-meter generation if a node's co-ordinates are negative.

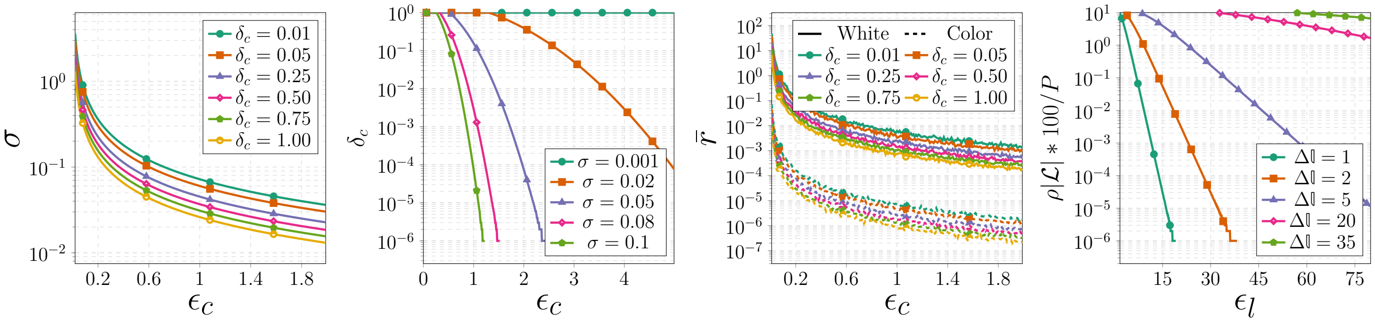

Fig. 6: DP performance for Gaussian Mixtures.

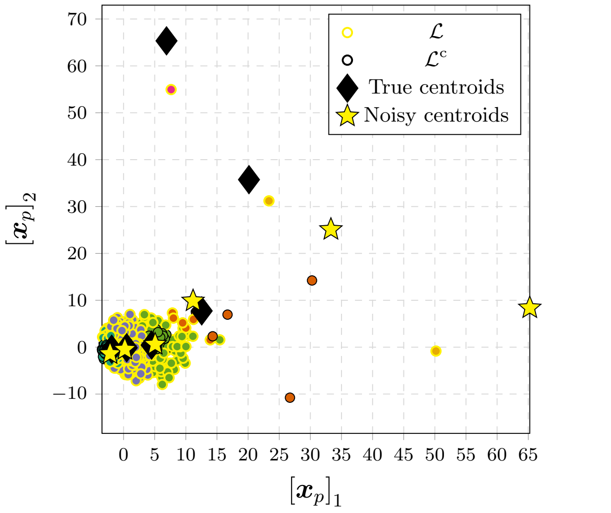

Fig. 7: Scatter plots of AMI dataset mapped in 2-d using MDS. The high sensitivity $\Delta_{\mathbb{C}} = 39.20$ is due to outliers.

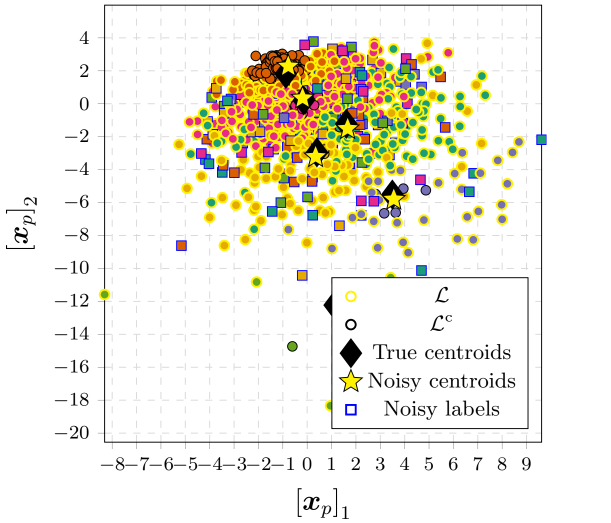

Fig. 8: Scatter plot of the AMI Dataset with outliers omitted, which has a sensitivity $\Delta_{\mathbb{C}} = 2.13$. Note that the points, which are originally 25d, are embedded in a 2d-space using MDS. $\epsilon_c = 30, \delta_c = 0.2, \epsilon_\ell = 30, \delta_\ell = 0$.

each node belongs to and depending on the application for which the data is being utilized, it is prudent that we analyze nodes that belong to similar classes and exclude class outliers. For example, in order to incentivize desired behavior by their customers, utilities might compare consumers with similar contracts, connected to distribution circuits in similar areas and climates and not mix commercial and residential customers. The clustering problem of the reduced dataset with outliers (commercial customers on closer inspection) removed, now with $P = 1409$, has a sensitivity equal to 2.13, which has reduced by a factor close to 20. We show the scatter plot of the dataset (using MDS) in fig. 8 with the true and noisy cluster centroids, where a colored Gaussian noise with $\epsilon_c = 30, \delta_c = 0.2, \epsilon_\ell = 30, \delta_\ell = 0$. In addition, the points whose labels were modified are also indicated using a □ marker. We show the (ϵ_c, δ_c) tradeoff in fig. 9 (left), and in fig. 9 (center), we compare our method against the white noise mechanism and the ones discussed by Balcan [17], Yu [16], and Ni [21]. It shows a clear improvement in the DP accuracy loss for a given (ϵ_c, δ_c) pair for our method compared to the others. The method proposed in [16] first eliminates the outliers as a preprocessing step, similar to our preprocessing step above, but differs in the noise mechanism used to perturb the centroids, and does not provide privacy guarantees on published labels. In fact, the 20-nearest neighbors based outlier elimination heuristic in [16] removes the same data points that we removed from our original dataset. Note that the DP schemes that utilize a Laplacian noise mechanism all have ϵ-DP guarantees, while our scheme has an (ϵ, δ)-DP guarantee. In essence, with a small increase δ above 0, our mechanism provides a far better DP accuracy loss performance.

Synthetic Load Profile Generation: Now, we use algorithm 2 to generate synthetic load profiles for this dataset. In fig. 10, we notice that the histograms of samples at each time interval are heavy tailed for every cluster except clusters 2 and 4. Note however that the number of nodes in these two clusters are significantly lower than the number of time intervals. We fit each time interval at each cluster to a group of heavy tailed distribution, and according to the Bayesian Information Criterion (BIC), the log-normal distribution is the best fit with the lowest BIC, for all clusters except 2 and 4. Since the number of samples in these two clusters are significantly lower than the number of time intervals (25), fitting any multivariate distribution becomes extremely inaccurate. Finally, in fig. 11, we show (in gray) 15 log-normal sample time series for each cluster with $\alpha = 15kW$. Here, the following privacy parameters were utilized: $(\epsilon_c, \delta_c, \epsilon_\ell, \delta_\ell) = (30, 0.2, 30, 0)$.

Standard Test Systems: The synthetic load profiles gener-

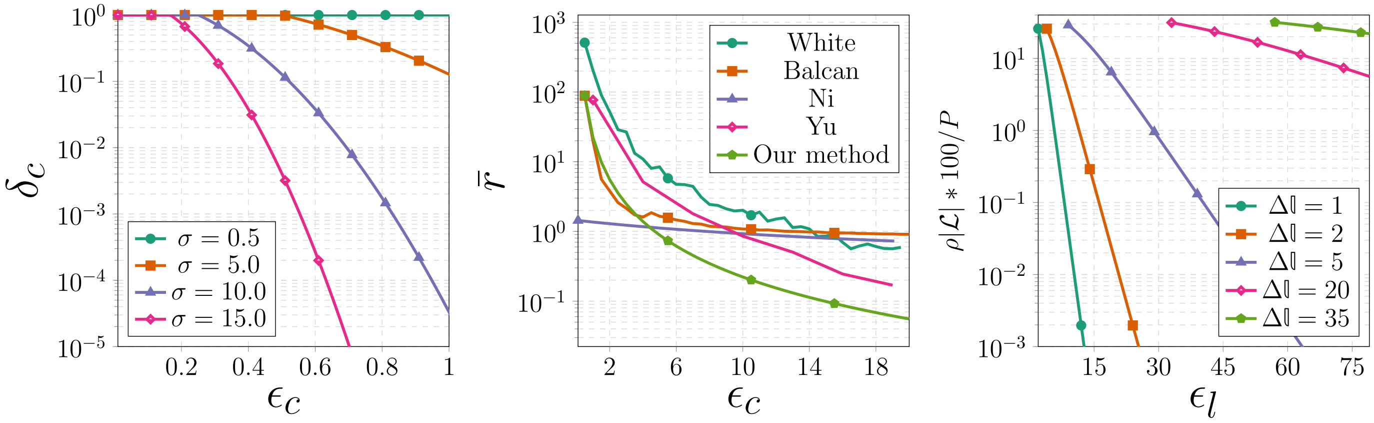

Fig. 9: Performance guarantees for the AMI Dataset with outliers omitted, which has a sensitivity of $\Delta \mathbb{c} = 2.13$. (Left) A plot of ϵ_c vs δ_c, (center) a plot of ϵ_c vs DP accuracy loss ($\delta_c = 0.01$), where our colored Gaussian mechanism with label noise is compared with the methods proposed by Balcan et al. [17], Yu et al. [16], Ni et al. [21] and the white Gaussian noise mechanism, and (right) a plot of ϵ_ℓ vs Expected percent of error in labels.

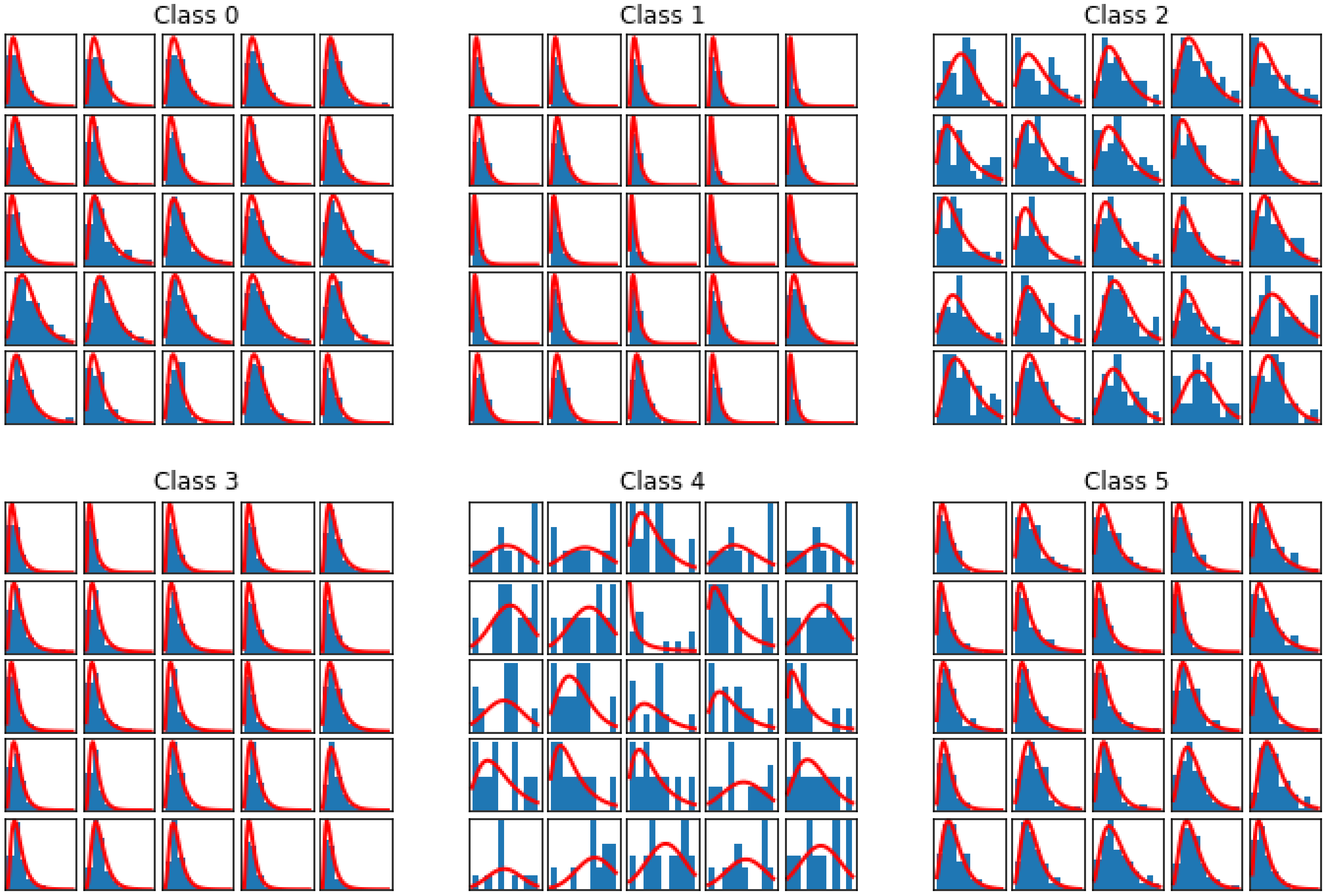

Fig. 10: Histograms of the load profiles at each time interval, grouped by cluster. The red curves show the best fit log-normal PDF.

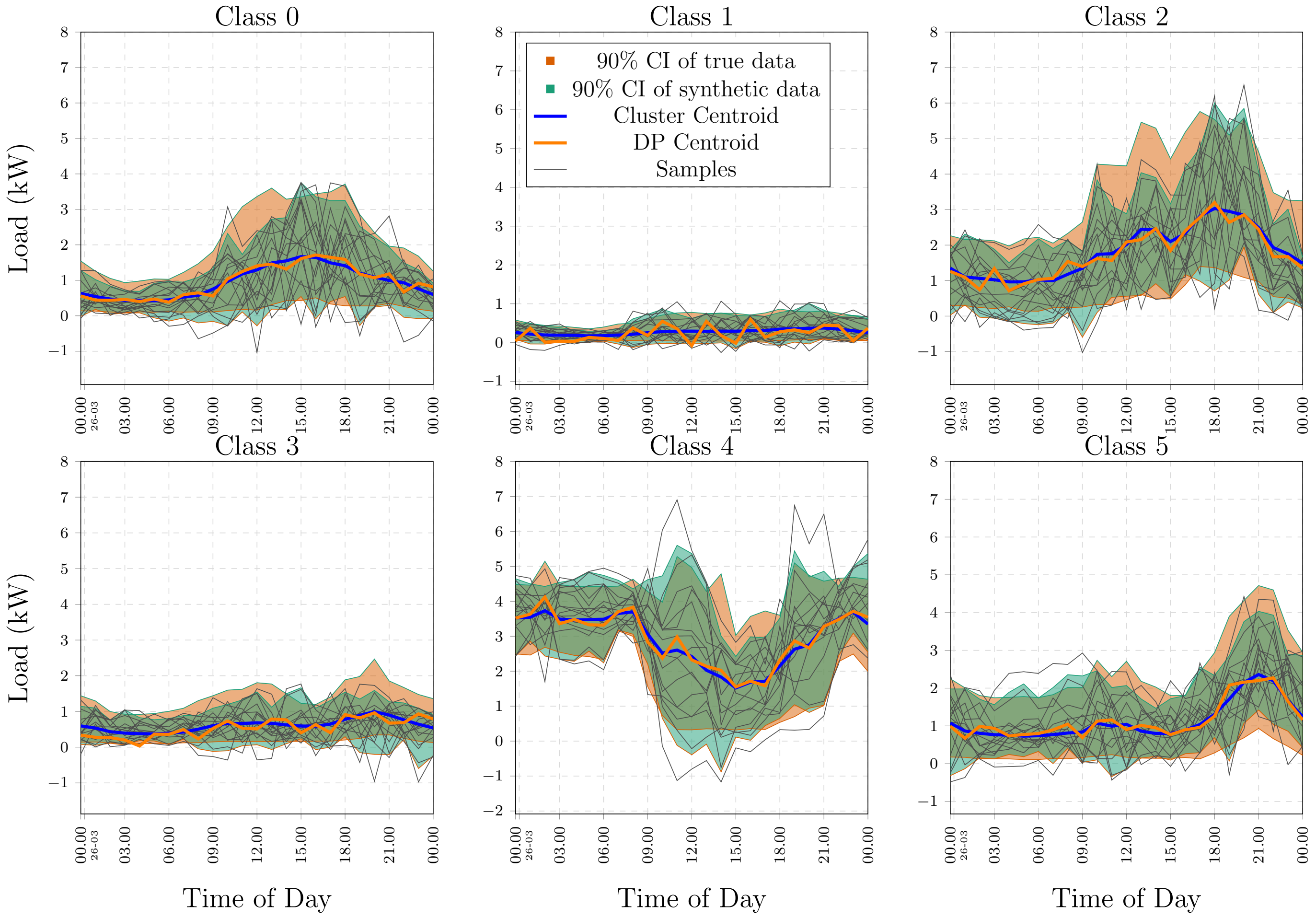

Fig. 11: DP Daily Load Shapes of the six clusters without outliers: $\epsilon_c = 30, \delta_c = 0.2$. The brown and green shaded regions indicate the 90% confidence interval (CI) of the true and the synthetically generated data points, respectively. These indicate the region between the 0.05 and 0.95 quantiles of the true dataset and that of the fitted log-normal distribution. The synthetic log-normal Daily Load Shapes of the six clusters are shown in gray lines.

ated in the previous section are now tested on two standard test cases, namely the MATPOWER 141-bus radial distribution system from [35] with $P' = 141$ and a modified balanced IEEE-123 test case from [36] with $P' = 123$. The testing methodology is as follows:

1) Randomly sample P' houses from the original dataset of $P = 1409$ houses, where the six classes are weighed according to their population. Let m_k be the size of class k in the sampled dataset.
2) Generate m_k samples using Algorithm 2 for all classes $k \in [K]$. Now we have $\boldsymbol{X}^{(t)} \in \mathbb{R}^{P' \times d}$ consisting of P' true load profiles for d time intervals and a corresponding $\boldsymbol{X}^{(s)} \in \mathbb{R}^{P' \times d}$ for the P' synthetic load profiles.
3) Load a test case with P' buses and set $X_{p,t}^{(t)}$ as the active power load for bus p, for all $p \in [P']$. Similarly, the reactive power load is set as 10% of the active power load.
4) Run an optimum power flow for this case with the modified loads and collect the voltage magnitude and phase information at each bus.
5) Repeat steps 3 and 4 for all time intervals $t \in [d]$.
6) Repeat steps 3, 4, and 5 by using $\boldsymbol{X}^{(s)}$ instead of $\boldsymbol{X}^{(t)}$.

In fig. 12 and fig. 13, we show the histograms of voltage magnitude and phase obtained under both cases with true and synthetic load profiles. The voltage magnitude and phase obtained for the synthetic load profiles provide a good match for those obtained for the true load profiles.

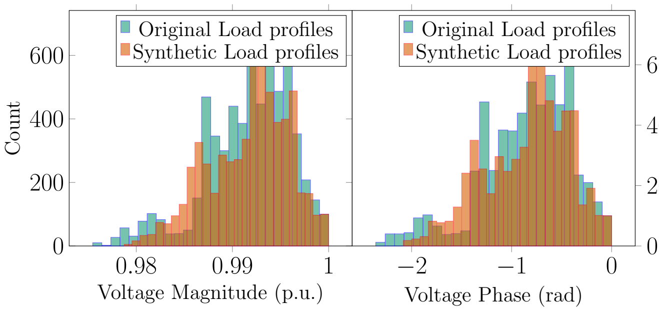

Fig. 12: Histogram of the voltage magnitude and phase under true and synthetic load profiles for the IEEE 123-bus case.

VI. Conclusion

In this paper, we presented an efficient, differentially private mechanism to answer summary statistics about data pertaining to a smart grid. To answer queries about the users in a dataset,

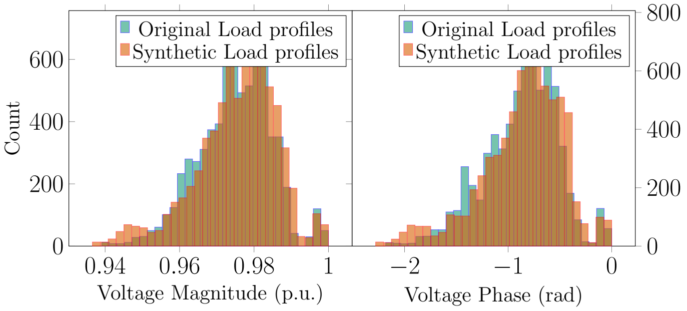

Fig. 13: Histogram of the voltage magnitude and phase under true and synthetic load profiles for the MATPOWER 141-bus case.

such as daily load shapes, we showed the use of clustering and the publication of the extracted clustered centroids rather than publishing individual daily loads. Our algorithm includes a colored Gaussian noise mechanism to guarantee differential privacy about the cluster centroids, and a novel discrete noise mechanism to guarantee differential privacy of the cluster labels. We also demonstrated the utility of our proposed mechanism using numerical simulations by answering queries such as daily load shapes and load duration curves of the houses in a power systems daily load dataset consisting of 1416 houses. In addition, we showed the importance of domain-specific knowledge to improve the utility of differential privacy. Finally, using the proposed clustering mechanism, we provided a mechanism to generate prototypical daily load shapes for the houses in a dataset.

Appendix A
Proof of Theorem 4

The joint probability mass function of the noise samples (stacked in a vector $\boldsymbol{\nu}$) can be written as:

$$f(\boldsymbol{\nu}) = (1-\rho)^{\gamma - \|\boldsymbol{\nu}\|_0} \left(\frac{\rho}{K-1}\right)^{\|\boldsymbol{\nu}\|_0},$$

where $\|\boldsymbol{a}\|_0$ is the zero "norm" operator that counts the number of non-zero elements in the vector $\boldsymbol{a}$. Consider the privacy leakage function:

$$L_{XX'}(\tilde{\boldsymbol{\ell}}) = \log \frac{f(\tilde{\mathbb{l}}(X)|X)}{f(\tilde{\mathbb{l}}(X')|X')} = \log \frac{f(\boldsymbol{\nu})}{f(\boldsymbol{\nu} + \mathbb{l}_{\mathcal{L}}(X') - \mathbb{l}_{\mathcal{L}}(X))},$$

where $\mathbb{l}_{\mathcal{L}}(X)$ and $\mathbb{l}_{\mathcal{L}}(X')$ map the points in $X \cap \mathcal{L}$ and $X' \cap \mathcal{L}$, respectively, to their labels. Letting $\mathbb{l}_{XX'} = \mathbb{l}_{\mathcal{L}}(X') - \mathbb{l}_{\mathcal{L}}(X)$, we have:

$$\begin{aligned} L_{XX'}(\tilde{\boldsymbol{\ell}}) &= \log\left((\rho^{-1}-1)(K-1)\right)^{\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0} \\ &= (\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0) \log\left((\rho^{-1}-1)(K-1)\right). \end{aligned}$$

For $K \geq 2$ and $\rho < 0.5$, the term $\log\left((\rho^{-1}-1)(K-1)\right)$ is non-negative. Thus, the probability that the absolute value of the privacy leakage function exceeds ϵ_ℓ is:

$$Pr(L_{XX'}(\tilde{\boldsymbol{\ell}}) \geq \epsilon_\ell) = Pr(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0 \geq \ell), \tag{21}$$

where $\ell := \epsilon_\ell / \log\left((\rho^{-1}-1)(K-1)\right)$.

a) Case 1: When $\ell > \|\mathbb{l}_{XX'}\|_0$, the above probability is 0. Furthermore, with $\Delta\mathbb{l} := \sup_{X'} \|\mathbb{l}_{XX'}\|_0$, we have:

$$\sup_{X'} Pr(L_{XX'}(\tilde{\boldsymbol{\ell}}) \geq \epsilon_\ell) = 0 \quad \text{if} \quad \ell > \Delta\mathbb{l}. \tag{22}$$

b) Case 2: When $\ell \leq \|\mathbb{l}_{XX'}\|_0$, the calculation of the probability in eq. (21) is not a straightforward task and involves numerical estimation.

Let M_0 and M_c be the number of zero elements and the number of elements which satisfy $\nu_p + \mathbb{l}_{XX',p} = 0$,[3] respectively, in $\boldsymbol{\nu}$. Based on the values of ν_p and $\mathbb{l}_{XX',p} \neq 0$,[4] we can make the following observations:

1) Whenever $\nu_p = 0$, the value of $(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0)$ increases by 1.
2) Whenever $\nu_p \neq 0$ such that $\nu_p + \mathbb{l}_{XX',p} = 0$, the value of $(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0)$ decreases by 1.
3) For the remaining possible value that ν_p can take, the value of $(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0)$ does not change. The number of such elements in $\boldsymbol{\nu}$ is $\|\mathbb{l}_{XX'}\|_0 - M_0 - M_c$.

From their respective definitions, it is clear that M_0 and M_c are jointly multinomial, i.e.,

$$\begin{aligned} &Pr(M_0 = m_0, M_c = m_c) \\ &= \frac{(\|\mathbb{l}_{XX'}\|_0)!(1-\rho)^{m_0}(\frac{\rho}{K-1})^{m_c}(\frac{\rho(K-2)}{K-1})^{(\|\mathbb{l}_{XX'}\|_0 - m_0 - m_c)}}{m_0!m_c!(\|\mathbb{l}_{XX'}\|_0 - m_0 - m_c)!}, \end{aligned} \tag{23}$$

and it can also be shown that $\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0 = M_0 - M_c$. When $\ell \leq \|\mathbb{l}_{XX'}\|_0$, eq. (21) can be simplified to:

$$\begin{aligned} &Pr(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0 \geq \ell) \\ &= \sum_{m_c=0}^{\|\mathbb{l}_{XX'}\|_0} \sum_{m_0=\ell+m_c}^{\|\mathbb{l}_{XX'}\|_0} Pr(M_0 = m_0, M_c = m_c) \\ &= (\|\mathbb{l}_{XX'}\|_0)! \left[\frac{\rho(K-2)}{K-1}\right]^{\|\mathbb{l}_{XX'}\|_0} \times \\ &\sum_{m_c=0}^{\|\mathbb{l}_{XX'}\|_0} \sum_{m_0=\ell+m_c}^{\|\mathbb{l}_{XX'}\|_0} \frac{\left[\frac{(K-1)(1-\rho)}{\rho}\right]^{m_0} (K-2)^{-(m_0+m_c)}}{m_0!m_c!(\|\mathbb{l}_{XX'}\|_0 - m_0 - m_c)!} \end{aligned} \tag{24}$$

Since the expression eq. (24) grows monotonically with $\|\mathbb{l}_{XX'}\|_0$, we can define the sensitivity as $\Delta\mathbb{l} := \sup_{X'} \|\mathbb{l}_{XX'}\|_0$. Thus,

$$\begin{aligned} &\sup_{X'} Pr(L_{XX'}(\tilde{\boldsymbol{\ell}}) \geq \epsilon_\ell) = \Delta\mathbb{l}! \left[\frac{\rho(K-2)}{K-1}\right]^{\Delta\mathbb{l}} \times \\ &\sum_{m_c=0}^{\Delta\mathbb{l}} \sum_{m_0=\ell+m_c}^{\Delta\mathbb{l}} \frac{\left[\frac{(K-1)(1-\rho)}{\rho}\right]^{m_0} (K-2)^{-(m_0+m_c)}}{m_0!m_c!(\Delta\mathbb{l} - m_0 - m_c)!} \end{aligned} \tag{25}$$

From eq. (22) and eq. (25), the proposed mechanism is $(\epsilon_\ell, \delta_\ell)$−PDP and thus, it is also $(\epsilon_\ell, \delta_\ell)$−DP from theorem 1.

[3] Note that we are using modulo K addition here.

[4] In case of $\mathbb{l}_{XX',p} = 0$, the value of $(\|\boldsymbol{\nu}+\mathbb{l}_{XX'}\|_0 - \|\boldsymbol{\nu}\|_0)$ does not change for any value of ν_p.

References

[1] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big Data Analytics for Dynamic Energy Management in Smart Grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, 2015.

[2] Y. Zhang, T. Huang, and E. F. Bompard, "Big Data Analytics in Smart Grids: A Review," *Energy informatics*, vol. 1, no. 1, pp. 1–24, 2018.

[3] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[4] S.-l. Yang, C. Shen *et al.*, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, 2013.

[5] S. Ramos, J. Duarte, J. Soares, Z. Vale, and F. J. Duarte, "Typical Load Profiles in the Smart Grid Context – A Clustering Methods Comparison," in *2012 IEEE Power and Energy Society General Meeting*. IEEE, 2012, pp. 1–8.

[6] D. D. Sharma and S. Singh, "Electrical Load Profile Analysis and Peak Load Assessment Using Clustering Technique," in *2014 IEEE PES General Meeting Conference & Exposition*, 2014.

[7] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, "Smart Meter Data Privacy: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2820–2835, 2017.

[8] S. Ruj and A. Nayak, "A Decentralized Security Framework for Data Aggregation and Access Control in Smart Grids," *IEEE transactions on smart grid*, vol. 4, no. 1, pp. 196–205, 2013.

[9] C. Efthymiou and G. Kalogridis, "Smart Grid Privacy via Anonymization of Smart Metering Data," in *2010 first IEEE international conference on smart grid communications*. IEEE, 2010, pp. 238–243.

[10] A. Narayanan and E. W. Felten, "No Silver Bullet: De-identification Still Doesn't Work," http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf, July 9, 2014.

[11] M. Barbaro, T. Zeller, and S. Hansell, "A face is exposed for AOL searcher no. 4417749," *New York Times*, vol. 9, no. 2008, p. 8, 2006.

[12] A. Narayanan and V. Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," in *29th IEEE Symposium on Security and Privacy*, May 2008.

[13] L. Sweeney, A. Abu, and J. Winn, "Identifying Participants in the Personal Genome Project by Name," *Available at SSRN 2257732*, 2013.

[14] P. U. C. of the State of Colorado, "Decision No. R11-0922," *Proposed Rules Relating to Smart Grid Data Privacy for Electric Utilities*, 2011.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of cryptography conference*, 2006, pp. 265–284.

[16] Q. Yu, Y. Luo, C. Chen, and X. Ding, "Outlier-Eliminated K-Means Clustering Algorithm Based on Differential Privacy Preservation," *Applied Intelligence*, vol. 45, no. 4, pp. 1179–1191, 2016.

[17] M.-F. Balcan, T. Dick, Y. Liang, W. Mou, and H. Zhang, "Differentially Private Clustering in High-Dimensional Euclidean Spaces," in *International Conference on Machine Learning*. PMLR, 2017, pp. 322–331.

[18] J. Ren, J. Xiong, Z. Yao, R. Ma, and M. Lin, "Dplk-means: A novel differential privacy k-means mechanism," in *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2017, pp. 133–139.

[19] C. Xia, J. Hua, W. Tong, and S. Zhong, "Distributed K-Means Clustering Guaranteeing Local Differential Privacy," *Computers & Security*, vol. 90, p. 101699, 2020.

[20] Z. Lu and H. Shen, "Differentially Private K-Means Clustering With Guaranteed Convergence," *arXiv preprint arXiv:2002.01043*, 2020.

[21] 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.

[22] A. Pinceti, O. Kosut, and L. Sankar, "Data-Driven Generation of Synthetic Load Datasets Preserving Spatio-Temporal Features," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, 2019.

[23] S. El Kababji and P. Srikantha, "A Data-Driven Approach for Generating Synthetic Load Patterns and Usage Habits," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4984–4995, 2020.

[24] M. N. Fekri, A. M. Ghosh, and K. Grolinger, "Generating Energy Data for Machine Learning With Recurrent Generative Adversarial Networks," *Energies*, vol. 13, no. 1, p. 130, 2020.

[25] A. Pinceti, L. Sankar, and O. Kosut, "Synthetic Time-Series Load Data via Conditional Generative Adversarial Networks," *arXiv preprint arXiv:2107.03545*, 2021.

[26] N. Ravi, A. Scaglione, and S. Peisert, "Colored Noise Mechanism for Differentially Private Clustering," *arXiv preprint arXiv:2111.07850*, 2021.

[27] X. Lou, R. Tan, D. K. Yau, and P. Cheng, "Cost of Differential Privacy in Demand Reporting for Smart Grid Economic Dispatch," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[28] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data, Ourselves: Privacy Via Distributed Noise Generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[29] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *2008 IEEE 24th international conference on data engineering*. IEEE, 2008, pp. 277–286.

[30] D. McClure, "Relaxations of Differential Privacy and Risk/Utility Evaluations of Synthetic Data and Fidelity Measures," Ph.D. dissertation, Duke University, 2015.

[31] C. Dwork, A. Roth *et al.*, "The Algorithmic Foundations of Differential Privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[32] W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, pp. 669–672, 1990.

[33] W. Jiang, C. Xie, and Z. Zhang, "Wishart Mechanism for Differentially Private Principal Components Analysis," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[34] A. Mead, "Review of the Development of Multidimensional Scaling Methods," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 41, no. 1, pp. 27–39, 1992.

[35] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2010.

[36] Y. Chai, L. Guo, C. Wang, Z. Zhao, X. Du, and J. Pan, "Network partition and voltage coordination control for distribution networks with high penetration of distributed pv units," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3396–3407, 2018.