Differentially Private \( K \)-Means Clustering Applied to Meter Data Analysis and Synthesis

Nikhil Ravi, Graduate Student Member, IEEE, Anna Scaglione, Fellow, IEEE, Sachin Kadam, Reinhard Gentz, Sean Peisert, Senior Member, IEEE, Brent Lunghino, Emmanuel Levijarvi, and Aram Shumavon

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.

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每日 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 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 stake-holders 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 \( x_i \) denotes the \( i \)-th element of a vector \( x \) (the \( j \)-th entry of a matrix \( X \) 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 \( x_p, p \in [P] \) embedded in \( \mathbb{R}^d \) that are in a database \( \mathcal{X} \), which we can organize as a \( P \times d \) matrix \( X := [x_1, \ldots, x_P]^\top \). Specifically, in an energy system that is considered in this paper, \( \mathcal{X} \) is the set of homes that a utility company serves, \( x_p \) 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 \( q(X) \) the function that maps \( \mathcal{X} \) onto the query answer, and denote the outcome by \( q \in \mathcal{Q} \), 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:

\[
q(X) = \frac{1}{|P|} \sum_{x \in \mathcal{X}} x, \tag{1}
\]

where \( P \geq 15 \) and \( \forall x \in \mathcal{X}, x \leq 0.15 q(X) \). 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 \( x \in \mathcal{X} \). Let us now consider an adversary who has queried for the average of the dataset \( \mathcal{X} \) and a neighboring dataset \( \mathcal{X}' \), which differ from \( \mathcal{X} \) by just one point \( x_0 \in \mathcal{X} \), i.e., \( \mathcal{X}' = \mathcal{X} \setminus \{x_0\} \). If both \( q(X) \) and \( q(X') \) are revealed to them, then by simple algebra:

\[
x_0 = Pq(X) - (P - 1)q(X'). \tag{2}
\]

Thus, they are able to infer the value of the point \( x_0 \in \mathcal{X} \) 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 \( q(X) \) the utility’s data, and receive a differentially privatized answer \( \tilde{q}(X) \) instead of the true query response \( q(X) \), as shown in Fig. 2. The quality of the DP response depends on the budget \((\epsilon_\ell, \delta_\ell, (\epsilon_i, \delta_i))\) 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.

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 \( X \) 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 \( q : \mathbb{R}^{P \times d} \rightarrow \{K\}^P \times \mathbb{R}^{K \times d} \). The problem can formally be posed as an optimization problem of the form:

\[
\arg \min_{\epsilon} \frac{1}{P} \sum_{k \in [K]} \sum_{p \in C_k} \|x_p - \epsilon_k\|_p, \tag{3a}
\]

\[
s.t. \bigcap_{k \in [K]} C_k = [P], \tag{3b}
\]

\(^1\)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.
where \( \mathcal{C} = \{ C_1, \ldots, C_k \} \) is the partition in \( K \) clusters and \( c_k \in \mathbb{R}^d \) is the centroid of a cluster \( C_k \), obtained by averaging the points \( \{ x_p \}_{p \in C_k} \). The objective in eq. (3a) minimizes the cost of clustering assignment \( \mathcal{C} \) 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:

\[
e_q(X) := \{ e(X), l(X) \},
\]

where

1) \( e(X) \) with outcome \( e = [ c_1^T, \ldots, c_{k-1}^T ]^T \) is the sub-query that returns the set of \( K \) centroids \( \{ c_k \in \mathbb{R}^d \}_{k \in [K]} \).

2) \( l(X) \) with outcome \( l \) is the sub-query that assigns a label to each feature point \( x_p \) that corresponds to an index of such centroids.

The performance of the \( K \)-means query outcome, \( q = [ e, l ] \), is measured in terms of the clustering loss which is defined as:

**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.,

\[
r(X, q) := \frac{1}{p} \sum_{p \in [p]} \| x_p - c_{l_p} \|^2_f, \quad \text{where} \quad q = [ e, l ]
\]

We note that this is equivalent to the minimizer of the clustering loss if the outcome of the \( K \)-means clustering algorithms was \( q \). 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 \( q \) and a DP answer \( \tilde{q} \), the DP query accuracy loss, or simply the DP accuracy loss, is given by:

\[
\tilde{r}(X, q, \tilde{q}) = \frac{r(X, q) - r(X, q)}{r(X, q)},
\]

where \( r(X, q, \tilde{q}) \) is the clustering loss if the outcome of the \( K \)-means clustering algorithms was \( \tilde{q} \). Since \( q \) is the minimizer of eq. (3), \( \tilde{r} \geq 0 \), 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 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 \( \tilde{q}(X) \), and has a random output \( \tilde{q} \in \mathbb{Q} \), with distribution \( f(\tilde{q}|X) \) (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 [(\( \epsilon, \delta \))-Differential Privacy]:** A randomized mechanism \( \tilde{q} \) is \( (\epsilon, \delta) \)-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

\[
\forall S, \quad \Pr(\tilde{q}(X) \in S) \leq \exp(\epsilon)\Pr(\tilde{q}(X') \in S) + \delta,
\]

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

Note that, since \( \delta \) is a bound that may not be tight, smaller values of \( \delta \) are possible. Hence, \( (\epsilon, \delta) \) 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 \( q \) is the answer generated by the data \( X \) or the data \( X' \) that differ by one element. Mathematically:

\[
L_{XX'}(q) := \log \frac{f(q|X)}{f(q|X')}. \tag{8}
\]

A randomized mechanism \( \tilde{q}(X) \) is \((\epsilon, \delta)\)-differentially private for \( X \) if and only if:

\[
\sup_{X'} Pr(L_{XX'}(\tilde{q}) > \epsilon) \leq \delta. \tag{9}
\]

It can be shown that \((\epsilon, \delta)\)-PDP is a strictly stronger condition than \((\epsilon, \delta)\)-DP.

Theorem 1 (PDP Implies DP) [30]: If a randomized mechanism is \((\epsilon, \delta)\)-PDP, then it is also \((\epsilon, \delta)\)-DP, i.e.,

\((\epsilon, \delta)\) - PDP \(\Rightarrow\) \((\epsilon, \delta)\) - DP, but \((\epsilon, \delta)\) - DP \(\not\Rightarrow\) \((\epsilon, \delta)\) - PDP.

In dealing with a multidimensional answer \( \tilde{q} \) (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 \((\epsilon_i, \delta_i)\) results for the scalar independent mechanism employed on a global \((\epsilon, \delta)\) result, rather than \( L_{XX'}(\tilde{q}) = \sum_{j=1}^{n} L_{XX'}(\tilde{q}_j) \).

Lemma 1 (Sequential Composition) [31]: If \( n \) randomized algorithms \( M_i, i = 1, 2, \ldots, n \), are \((\epsilon_i, \delta_i)\)-DP, then the sequential execution of these algorithms on the database \( X \) provides \((\sum_{i=1}^{n} \epsilon_i, \sum_{i=1}^{n} \delta_i)\)-differential privacy.

In this paper, we use Definition 4 which has an immediate statistical interpretation: if values of \((\epsilon_i, \delta_i)\) 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 \( \tilde{q}(X) = [\tilde{c}(X), \tilde{I}(X)] \) that renders the clustering query answer \((\epsilon, \delta)\)-DP.

III. An \((\epsilon, \delta)\)-DP Mechanism for \( K \)-Means Clustering

The \( K \)-means clustering query, as described in Section II-B1, has two components: the centroids and the labels \( q(X) = [c(X), I(X)] \). 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

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:

\[
\tilde{c}(X) = c(X) + \eta = c + \eta, \tag{10}
\]

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

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

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

and \( \Delta c \) is the query sensitivity given by:

\[
\Delta c = \sup_{X'} \|c(X) - c(X')\|_2.
\]

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 Savitzky-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.
Thus, for all neighboring datasets \( X' \in \mathcal{X} \) of \( X \) that differ in one element, we have that:
\[
(\epsilon(X) - \epsilon(X'))^\top \Gamma(\epsilon(X) - \epsilon(X')) \leq \Delta \epsilon^2 \leq \gamma_c.
\]

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:
\[
\min_{\Gamma} \text{tr}(\Gamma^{-1})
\]
\[\text{s.t.} \quad (\epsilon(X) - \epsilon(X'))^\top \Gamma(\epsilon(X) - \epsilon(X')) \leq \gamma_c, \quad \forall X' \in \mathcal{X}. \tag{15}\]

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 \( \Gamma \))**: Let the matrix \( C_{XX'} \) contain as its columns all possible \((\epsilon(X) - \epsilon(X'))\), \(X' \in \mathcal{X}\) and let us assume that \( C_{XX'} \) is full row rank, and that the first \( Kd \) columns of \( C_{XX'} \) corresponding to the set \( D \subseteq \mathcal{X} \) have the smallest norms and are linearly independent, forming the matrix we refer to as \( \lambda_{XX'}^* \). Then, the optimization problem in eq. (15) has a unique solution and it evaluates to:
\[
\Gamma^* = R_{XX'}^{-1}, \tag{16}\]
where \( \lambda^* \) 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 \( D \), and:
\[
R_{XX'} := \sum_{X \in D} \lambda_{XX'}(\epsilon(X) - \epsilon(X'))(\epsilon(X) - \epsilon(X'))^\top, \tag{17}\]
where \( \lambda_{XX'} \), \( \forall X' \in D \) are the non-zero Lagrange multipliers for the problem in eq. (15). Their values are:
\[
\lambda^*_j = v_j^{-2}, \quad \text{where} \quad v = \gamma_c M^{-1}1, \quad M_{ij} = \left[ C_{XX'}^{-1} \right]_{ij}. \tag{18}\]

Thus, the mechanism in eq. (12) is \((\epsilon_c, \delta_c)-DP\) 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.
remains in the permissible set \( K \) is the modulo \( K \) in the DP Label mechanism. Fig. 4. The composition of the set of points whose labels are treated with RA VI each of the points in the modular addition of i.i.d noise samples to the labels of in the Appendix.

where \( K \) might flip on the removal of a point from the dataset. Thus, 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 \( L \subseteq [P] \), which is the set of all the points whose labels change if any other point is removed. The composition of \( L \) is illustrated in Fig. 4, where we have a set of points \( \{0, \ldots, 19\} \). Here, when 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 \( L \).

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:

\[
\tilde{I}(X)_p = \begin{cases} \{I(X)_p \} \oplus_K v_p, & p \in L \\ \{I(X)_p\}, & \text{otherwise}, \end{cases}
\]

where \( \{v_p\}_{p \in L} \) 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 \( L \). In the following theorem, we provide the DP guarantees afforded by this mechanism with the proof in the Appendix.

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

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

where \( K \geq 2, \rho < 0.5 \), \( \Delta \) is the query sensitivity given by:

\[
\Delta = \sup_{X'} \|I(X') - I(X)\|_0.
\]

\[
\delta_\ell = \Delta \left[ \frac{(K-2)}{K-1} \right]^{\Delta} \times \sum_{m_0=0}^{\Delta} \sum_{m_\ell=0}^{\Delta} \text{mc}(\Delta - m_0 - m_\ell)!.
\]

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

\[
E_{\rho} \left[ \sum_{p \in L} 1_{\{v_p \neq 0\}} \right] = |L|/\rho.
\]

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

\[
r(X, \tilde{\ell}) := \frac{1}{\rho} \sum_{p \in [P]} \|x_p - \tilde{\ell}_p\|^2_2.
\]

**Algorithm 1 DP Mechanism With Colored Gaussian Noise and Randomized Labels**

1: **procedure** DPK-MEANS(\( \epsilon, \ell, \delta, \epsilon_\ell, \delta_\ell \))
2: \( q(\epsilon, \ell) \leftarrow \text{Cluster } X \text{ into } K \text{ clusters}.
3: Calculate \( \Gamma \) according to eq. (16) with \( \epsilon_\ell \) and \( \delta_\ell \).
4: \( \tilde{c} \leftarrow c + \eta, \text{ where } \eta \sim N(0, \Gamma). \)
5: \( \tilde{\ell} \leftarrow \ell \oplus v, \text{ with } v \text{ distributed according to the conditions of theorem 4}. \)
6: **end procedure**

**Output:** \( \tilde{q} = (\tilde{c}, \tilde{\ell}). \)

and \( \delta_\ell \) is 0 when \( \epsilon_\ell > \Delta \times \log((1 - \rho)/(K - 1)) \) and when \( \epsilon_\ell < \Delta \times \log((1 - \rho)/(K - 1)) \), it is given by

\[
\delta_\ell = \Delta \left[ \frac{\rho(K - 2)}{K - 1} \right]^{\Delta/2} \times \sum_{m_0=0}^{\Delta} \sum_{m_\ell=0}^{\Delta} mc(\Delta - m_0 - m_\ell)!.
\]

C. Mechanism and Its Differential Privacy Guarantees

In Sections III-A, 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).

**Corollary 1 [Clustering Query is (\( \epsilon, \delta \))-DP]:** For any two neighboring datasets \( X \) and \( X' \), the mechanisms described in eq. (10) and eq. (19) together provide \( (\epsilon, \delta)\)-DP guarantees with:

\[
\epsilon = \epsilon_\ell + \epsilon, \quad \delta = \delta_\ell + \delta_\ell.
\]

The proof follows from Theorems 3, 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...
η and ensure that the argument of the log operation remains positive, summarized in algorithm 2, where the exponential of the generated samples. This approach is to generate multivariate Gaussian samples, before finally outputting private. To obtain a DP covariance, we apply the Wishart element in element is the exponent (resp. logarithm) of the corresponding 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. \( \log(X) \)) denote the matrix where each element is the exponent (resp. logarithm) of the corresponding element in \( X \), 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 \( \alpha \) parameter is used to ensure that the argument of the log operation remains positive, and \( \eta_k \) and \( \Phi_k \) are the noise added to the cluster centroid and to the covariance matrix of the data pertaining to the cluster \( k \).

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 \( \mathcal{N}(\mu_k, \sigma I) \) \( k \in [K] \) with \( \mu_k \in \mathbb{R}^2 \). Note that because the mixture components covariance is \( I \), 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 \( L \) 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 \( \delta \)'s, the variation of \( \delta \) for various noise variances and the DP accuracy loss for various \( \delta \)'s. As expected, in order to achieve a stronger privacy guarantee (i.e., lower \( \delta \) values for given \( \sigma \)), 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 \( \Delta \) increases, the expected percent of errors, \( \rho|L| * 100/P \), also increases for a given privacy budget.

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, \( x_p \), 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 (MDS) and, as

\[ \text{Algorithm 2 Algorithm to Generate Load Profiles} \]

1: procedure LOGNORMSAMPLES
2: Cluster \( X \) into \( K \) clusters.
3: for \( k \in [K] \) do
4: \( \hat{X}_k \leftarrow \log([x_{p_1}, \ldots, x_{p_{|C_k|}}]) + \alpha I^T \)
5: \( \hat{\mu}_k = \frac{1}{|C_k|} \hat{X}_k 1 + \eta_k \)
6: \( \hat{\Sigma}_k = \frac{1}{|C_k|} \hat{X}_k \hat{X}_k^T + \Phi_k \)
7: \( \hat{X} \leftarrow \text{samples drawn from } \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k) \)
8: \( \hat{X} \leftarrow \exp(\hat{X}) - \alpha I^T \).
9: end for
10: end procedure

2MDS 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.
before, in Fig. 7, the points in $L$ 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 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 $\Delta c = 2.13$. Note that the points, which are originally 25d, are embedded in a 2d-space using MDS. $\epsilon_c = 30, \delta_c = 0.2, \epsilon_l = 30, \delta_l = 0$.

In addition, the points whose labels were modified are also indicated using a □ marker. We show the ($\epsilon_c, \delta_c$) 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 et al. [17], Yu et al. [16], and Ni et al. [21]. It shows a clear improvement in the DP accuracy loss for a given ($\epsilon_c, \delta_c$) 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 $\epsilon$-DP guarantees, while our scheme has an ($\epsilon, \delta$)-DP guarantee. In essence, with a small increase $\delta$ 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
Fig. 9. Performance guarantees for the AMI Dataset with outliers omitted, which has a sensitivity of $\Delta c = 2.13$. (Left) A plot of $\epsilon_c$ vs $\delta_c$, (center) a plot of $\epsilon_c$ 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 $\epsilon_\ell$ 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.

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 $\alpha = 15$\text{KW}. 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 generated 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 $X^{(t)} \in \mathbb{R}^{P' \times d}$ consisting of $P'$ true load profiles for $d$ time intervals and a corresponding $X^{(s)} \in \mathbb{R}^{P' \times d}$ for the $P'$ synthetic load profiles.

3) Load a test case with $P'$ buses and set $X^{(t)}_{p,t}$ 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 $X^{(s)}$ instead of $X^{(t)}$.

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.

**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, 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**

**PROOF OF THEOREM 4**

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

$$f(\mathbf{v}) = (1 - \rho)^{\|\mathbf{a}\|_0} \left(\frac{\rho}{K-1}\right)^{\|\mathbf{v}\|_0},$$

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

$$L_{XX}(\hat{\ell}) = \log \frac{f(\mathbf{v} + \mathbb{I}_{XX} X|X)}{f(\mathbf{v} + \mathbb{I}_{XX}(X') - \mathbb{I}_{XX}(X))},$$

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

$$L_{XX}(\hat{\ell}) = \log \left(\frac{\rho}{K-1}\right)^{\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0}$$

$$= (\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0) \log \left(\frac{\rho}{K-1}\right)^{(K-1)}) .$$

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

$$Pr\left(L_{XX}(\hat{\ell}) \geq \epsilon \right) = Pr(\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0 \geq \hat{\ell}) .$$

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

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

$$\sup_{X'} Pr\left(L_{XX}(\hat{\ell}) \geq \epsilon \right) = 0 \text{ if } \hat{\ell} > \Delta \|.$$

b) Case 2: When $\ell \leq \|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0$, 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 $\mathbb{v}_p + \mathbb{I}_{XX} X|X\|_0 = 0$, respectively, in $\mathbb{v}_p$. Based on the values of $\mathbb{v}_p$ and $\mathbb{I}_{XX} X|X\|_0$, we can make the following observations:

1) Whenever $\mathbb{v}_p = 0$, the value of $\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0$ increases by 1.

2) Whenever $\mathbb{v}_p \neq 0$ such that $\mathbb{v}_p + \mathbb{I}_{XX} X|X\|_0 = 0$, the value of $\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0$ decreases by 1.

3) For the remaining possible value that $\mathbb{v}_p$ can take, the value of $\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0$ does not change. The number of such elements in $\mathbb{v}$ is $\|\mathbb{I}_{XX} X|X\|_0 - M_0 - M_c$.

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

$$Pr(M_0 = m_0, M_c = m_c) = (\|\mathbb{I}_{XX} X|X\|_0)^{m_0} (\mathbb{I}_{XX} X|X\|_0)^{m_c} \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)} \cdot \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)},$$

and it can also be shown that $\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0 = M_0 - M_c$.

When $\ell \leq \|\mathbb{I}_{XX} X|X\|_0$, the calculation of the probability in eq. (21) can be simplified to:

$$Pr(\|\mathbf{v} + \mathbb{I}_{XX} X|X\|_0 - \|\mathbf{v}\|_0 \geq \ell)$$

$$= \sum_{m_0 = 0}^{M_0} \sum_{m_c = 0}^{M_c} Pr(M_0 = m_0, M_c = m_c)$$

$$= (\|\mathbb{I}_{XX} X|X\|_0)^{m_0} (\mathbb{I}_{XX} X|X\|_0)^{m_c} \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)} \cdot \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)},$$

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

$$\sup_{X'} Pr\left(L_{XX}(\hat{\ell}) \geq \epsilon \right) = \Delta \| \left(\frac{\rho(K-2)}{K-1}\right)^{\|\mathbb{I}_{XX} X|X\|_0} \times \sum_{m_0 = 0}^{M_0} \sum_{m_c = 0}^{M_c} \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)} \cdot \left(\frac{\rho(K-2)}{K-1}\right)^{(\|\mathbb{I}_{XX} X|X\|_0 - m_0 - m_c)}.$$
REFERENCES

[1] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, “Big data analytics for dynamic energy management in smart grids,” Big Data Anal., 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 Inform., 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 Trans. Smart Grid, vol. 10, no. 3, pp. 3125–3148, May 2019.

[4] K.-L. Zhou, S.-L. Yang, and C. Shen, “A review of electric load classification in smart grid environment,” Renew. Sustain. Energy Rev., vol. 24, pp. 103–110, Aug. 2013.

[5] S. Ramos, J. M. M. Duarte, J. Soares, Z. Vale, and F. J. Duarte, “Typical load profiles in the smart grid context—A clustering methods comparison,” in Proc. IEEE Power Energy Soc. General Meeting, 2012, pp. 1–8.

[6] D. D. Sharma and S. N. Singh, “Electrical load profile analysis and peak load assessment using clustering technique,” in Proc. IEEE PES General Meeting Conf. Expo., 2014, pp. 1–5.

[7] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, “Smart meter data privacy: A survey,” IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2820–2835, 4th Quart., 2017.

[8] S. Ruj and A. Nayak, “A decentralized security framework for data aggregation and access control in smart grids;” IEEE Trans. Smart Grid, vol. 4, no. 4, pp. 196–205, Mar. 2013.

[9] C. Efthymiou and G. Kalogridis, “Smart grid privacy via anonymization of smart metering data,” in Proc. 1st IEEE Int. Conf. Intell. Conf. Intell. Grid Comput., 2010, pp. 238–243.

[10] A. Narayanan and E. W. Felten, “No Silver Bullet: De-Identification Still Doesn’t Work,” Sep. 2014. [Online]. Available: http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[11] M. Bartho, T. Zeller, and S. Hansell, A Face is Exposed for AOL Searcher No. 4417749, vol. 9, New York Times, New York, NY, USA, 2006, p. 8.

[12] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in Proc. 29th IEEE Symp. Security Privacy, May 2008, pp. 111–125.

[13] L. Sweeney, A. Abu, and J. Winn, “Identifying participants in the personal genome project by name,” Harvard Univ., Cambridge, MA, USA, White Paper 1021-1, Apr. 2013. [Online]. Available: http://dataprivacylab.org/projects/ppp/

[14] Proposed Rules Relating to Smart Grid Data Privacy for Electric Utilities, Decision No. R11-0922, Public Utility Commission State Colorado, Denver, CO, USA, 2011.

[15] D. J. Wu, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Proc. Theory Cryptogr. Conf., 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,” Appl. Intell., vol. 45, no. 4, pp. 1179–1191, 2016.

[17] M. P. Balcan, T. Dick, Y. Liang, W. Mou, and H. Zhang, “Differentially private clustering in high-dimensional Euclidean spaces,” in Proc. Int. Conf. Mach. Learn., 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 Proc. IEEE 2nd Int. Conf. Data Sci. Cybersp. (DSC), 2017, pp. 133–139.

[19] C. Xia, J. Hua, W. Tong, and S. Zhong, “Distributed K-means clustering guaranteeing local differential privacy,” Comput. Security, vol. 90, Mar. 2020, Art. no. 101699.

[20] Z. Lu and H. Shen, “Differentially private K-means clustering with guaranteed convergence,” 2020, arXiv:2002.01043.

[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, Feb. 2021.

[22] A. Pinceti, O. Kosut, and L. Sankar, “Data-driven generation of synthetic load datasets preserving spatio-temporal features,” in Proc. IEEE Power Energy Soc. General Meeting (PESGM), 2019, pp. 1–5.

[23] S. El Kababji and P. Srikantha, “A data-driven approach for generating synthetic load patterns and usage habits,” IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 4984–4995, Nov. 2020.

[24] A. Pinceti, L. Sankar, and O. Kosut, “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,” 2021, arXiv:2107.03545.

[26] N. Ravi, A. Scaglione, and S. Fei, “Colored noise mechanism for differentially private clustering,” 2021, arXiv:2111.07850.

[27] X. Lou, R. Tan, D. K. Y. Yau, and P. Cheng, “Cost of differential privacy in demand reporting for smart grid economic dispatch,” in Proc. IEEE INFOCOM Conf. Comput. Commun., 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 Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn., 2006, pp. 486–503.

[29] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, “Privacy: Theory meets practice on the map,” in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 277–286.

[30] D. McClure, “Relaxations of differential privacy and risk/utility evaluations of synthetic data and fidelity measures,” Ph.D. dissertation, Dept. Stat. Sci., Duke Univ., Durham, NC, USA, 2015.

[31] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014.

[32] W. H. Press and S. A. Teukolsky, “Savitzky-Golay smoothing filters,” Comput. Phys., vol. 4, no. 6, pp. 660–672, 1990.

[33] W. Jiang, C. Xie, and Z. Zhang, “Wishart mechanism for differentially private principal components analysis,” in Proc. AAAI Conf. Artif. Intell., vol. 30, 2016, pp. 1730–1736.

[34] A. Mead, “Review of the development of multidimensional scaling methods,” J. Roy. Stat. Soc. D, Stat., 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 Trans. Power Syst., vol. 26, no. 1, pp. 12–19, Feb. 2011.

[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 Trans. Power Syst., vol. 33, no. 3, pp. 3396–3407, May 2018.

Nikhil Ravi (Graduate Student Member, IEEE) received the B.E. degree in electronics and communication engineering from the PES Institute of Technology, India, in 2017, and the M.S. degree in electrical engineering from Arizona State University, USA, in 2021. He is currently pursuing the Ph.D. degree with Cornell Tech, Cornell University. His research interests lie in fault-tolerant and privacy-preserving algorithms for complex networked dynamical systems.

Anna Scaglione (Fellow, IEEE) is currently a Professor of Electrical and Computer with Cornell Tech, The New York City Campus of Cornell University. Before that, she held faculty positions with Arizona State University, the University of California at Davis, Cornell University (the first time), and the University of New Mexico. Her expertise and research considers theoretical and applied problems in statistical signal processing, communications theory, and cyber-physical systems for energy delivery. She received the 2013 IEEE Donald G. Fink Prize Paper Award, the 2000 IEEE Signal Processing Transactions Best Paper Award the NSF CAREER Grant in 2002, and the 2020 Technical Achievement Award from the IEEE Communication Society Technical Committee on Smart Grid Communications. She is a co-recipient with her students of several best student papers awards at conferences and received the 2013 IEEE Signal Processing Society Young Author Best Paper Award with one of her Ph.D. students. She was a Distinguished Lecturer of the Signal Processing Society in 2019 and 2020.
Sachin Kadam received the B.Eng. degree in electronics and communication engineering from the People’s Education Society Institute of Technology, Bengaluru, India, in 2007, the M.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT) Kanpur, India, in 2012, and the Ph.D. degree from IIT Bombay, India, in 2020. He is currently a Postdoctoral Researcher with Prof. Dong In Kim with Sungkyunkwan University, Suwon, Gyeonggi, South Korea. His research interests include the design and analysis of wireless and M2M networks, semantic communications, differential privacy, and learning. He was a recipient of Scholarship Foundation for Excellence, California, USA, during the B.Eng. degree.

Reinhard Gentz received the B.Sc. and M.Sc. degrees in electrical engineering from the Karlsruhe Institute of Technology, Germany, in 2010 and 2014, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, USA, in 2017. He is currently a Software Development Engineer, Networking with Amazon. His research interests include networking, cybersecurity, software-defined networks, and algorithm development.

Sean Peisert (Senior Member, IEEE) received the Ph.D. degree in computer science from UC San Diego. He leads the Computer Security Research and Development with Lawrence Berkeley National Laboratory, Berkeley, CA, USA, and is an Adjunct Professor with the University of California at Davis, Davis, CA, USA. He is the Editor-in-Chief of IEEE SECURITY AND PRIVACY. He is a Senior Member of the Association of Computing Machinery.

Brent Lunghino received the B.S. degree in geological sciences from Brown University and the M.S. degree in computational and mathematical engineering from Stanford University. He is currently a Senior Data Scientist with Kevala. His work focuses on using data-driven models to answer questions about how to integrate distributed energy resources into the electrical grid. He received the NSF Graduate Research Fellowship in 2015.

Emmanuel Levijarvi is currently the Vice President of Software and Platform with Kevala. He has spent over 20 years developing software systems at scale and leading software engineering teams in the development of these systems. Recently he has led the development of Kevala’s software platform, which analyzes the impact of distributed energy resources on the electrical grid.

Aram Shumavon received the A.B. degree in economics and public policy from the University of Chicago. He has more than 20 years of experience in regulated energy markets, with a focus on high penetration distributed and renewable energy resources. He is the Founder and the CEO of Kevala, Inc., and the Prior Founder and the Executive Director of the nonprofit Distributed Energy Consumer Advocates. He has helped develop state and federal policy, managed the development of regulatory strategies, and created innovative market infrastructure and policy analysis tools. His work includes more than a decade as a senior analyst with the California Public Utilities Commission, where he headed the state’s analysis of key wholesale electricity market design and environmental compliance efforts and led long-term procurement and planning efforts focused on California’s high renewables penetration future and renewables integration strategies.