Enhancing Privacy and Availability for Data Clustering in Intelligent Electrical Service of IoT

Jinbo Xiong, *Member, IEEE*, Jun Ren, Lei Chen, *Member, IEEE*, Zhiqiang Yao, Mingwei Lin, Dapeng Wu, *Senior Member, IEEE*, and Ben Niu

***Abstract*—The ever-growing demand for electrical energy of sensing devices in the Internet of Things (IoT) has led to generating large amounts of electricity consumption data. Electricity service providers often use wireless sensor networks to collect sensing devices' electricity consumption data for statistical analysis, so as to provide sensing devices with improved electrical services. As an important data mining technique, while data clustering excels in dealing with such massive data, it imposes the risk of privacy disclosure in the process of data clustering. In an effort of solving this problem, Blum *et al.* proposed a differential privacy *k*-means algorithm, effectively preventing privacy disclosure. However, the availability of data clustering results is reduced due to the data distortion in Blum's algorithm. In this paper, we propose a privacy and availability data clustering (PADC) scheme based on *k*-means algorithm and differential privacy, which enhances the selection of the initial center points and the distance calculation method from other points to center point. Moreover, PADC attempts to reduce the outlier effect through detecting outliers during the clustering process. Security analysis indicates that our scheme satisfies the goal of differential privacy and prevents privacy information disclosure. Meanwhile, performance evaluation shows that our scheme, at the same privacy level, improves the availability of clustering results compared to the existing differential privacy *k*-means algorithms, suggesting that our proposed PADC scheme outperforms others for intelligent electrical service in IoT.**

***Index Terms*—Data clustering, differential privacy, Internet of Things (IoT), *k*-means algorithm, privacy protection.**

Manuscript received January 29, 2018; revised April 28, 2018; accepted May 22, 2018. Date of publication June 1, 2018; date of current version May 8, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61502489, Grant 61402109, Grant 61370078, Grant 61502102, and Grant 61502103, and in part by the Chongqing Key Laboratory of Optical Communication and Networks Research Fund under Grant KLOCN2018001. An earlier version of this paper "DPLK-Means: A Novel Differential Privacy K-Means Mechanism" was presented at the Proceedings of the IEEE International Conference on DSC, Shenzhen, China, June 26–29, 2017. *(Corresponding author: Zhiqiang Yao.)*

J. Xiong, J. Ren, Z. Yao, and M. Lin are with the Fujian Provincial Key Laboratory of Network Security and Cryptology, College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China (e-mail: jbxiong@fjnu.edu.cn; jun_ren@outlook.com; yzqfzfj@fjnu.edu.cn; linmwcs@163.com).

L. Chen is with the College of Engineering and Computing, Georgia Southern University, Statesboro, GA 30461 USA (e-mail: lchen@georgiasouthern.edu).

D. Wu is with the Chongqing Key Laboratory of Optical Communication and Networks, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: wudapengphd@gmail.com).

B. Niu is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: niuben@iie.ac.cn).

Digital Object Identifier 10.1109/JIOT.2018.2842773

I. Introduction

AS AN important part of a new generation of information technology, the Internet of Things (IoT) [1] aims to enable the exchange of information among objects, and is widely used in the integration of networks through communication sensing technologies, such as intellisense, recognition technology, pervasive computing, and fog/edge computing [2], [3]. Therefore, the IoT is also known as the third wave of the information revolution. Advances in new technologies have prompted to the rapid growth of diversified electronic products, and subsequently led to a surge in electricity consumption, while generating large amounts of data that tie back to the electricity consumption.

With the concept of smart city [4], there has many agencies that use IoT in the design of smart grid, where the IoT technology may effectively integrate communications and power system infrastructure resources, provide information and communication services to the power system, and enhance the details of power system information as an important technical support for the smart grid. Furthermore, managers can use wireless sensor network (WSN) [5] and device-to-device communications [6]–[8], formed by multiple sensor nodes communicating with one another, to obtain large volumes of electricity data in real time. Under the IoT framework, the perception layer technology flexibly supports WSNs with various node scales, under which the electricity consumption data is uploaded to the network layer and further passed to the application layer for data integration and analysis. This hierarchy can effectively reduce transmission delays and support monitoring applications that require highly real-time performance. Accurate and meaningful statistical analyses of these electricity consumption data are extremely useful in providing more reliable and satisfied services to customers. Data clustering, an important data mining technique, particularly aims to assist in handling massive data [9]. It has been extensively studied and employed in statistics and artificial intelligence applications, including market research, pattern recognition, information retrieval, image processing, intrusion detection, among many others [10]. The principle of data clustering is to better comprehend the distribution and meaning of complex and massive data by dividing them into classes of the same or similar nature [11]. For example, data clustering can be used to analyze electricity consumption data for identifying the peaks and valleys of electricity consumption and the unique electrical

2327-4662 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

behaviors of sensing devices. Ultimately, by utilizing such results, electricity service providers can accurately identify sensitive devices and quantify the degree of sensitivity, thereby supporting targeted devices service strategy, controlling the cost of electricity labor and further enhancing corporate's public image. Security risks obviously exist as the data analysis from the IoT may contain sensitive or private information from devices and users [12]. It is possible and practical to encrypt data at the user end and secure the data exchanges between users and the data center, however, during the data clustering and analysis process at the data center, sensitive data such as users' house addresses and electricity usage can potentially be obtained by adversaries for malicious purposes.

Such security and privacy threats have imposed serious issues [13] and drawn attention from researchers in this area [14]. Proposed as a new research direction for data mining in [15], privacy protection data mining aims to protect the privacy information under the premise of obtaining effective mining results. In the existing methods under this research direction, the security multiparty computation has been proven to enhance the mining precision at the expense of efficiency. These works have focused on the association rules and classification algorithms, whereas the research on clustering of privacy protection is significantly insufficient, with focuses on centralized data and vertical distribution data [16]. As an example, Vaidya and Clifton [17] proposed a secure multi-k-means algorithm based on vertical distribution data, and Jha *et al.* [18] and Inan *et al.* [19] proposed a clustering algorithm for privacy protection of horizontal distribution data. In the security multiparty computation process, given that data analysts need to use the appropriate algorithm to eliminate the impact of encryption, caused by earlier data encryption performed by users and managers, on mining results, such approach typically suffers from low excavation efficiency due to high communication overheads and huge computation costs [20]. Wu *et al.* [21] combined key distribution with trust management to construct a novel dynamic trust relationships aware data privacy protection mechanism. Miao *et al.* [22] constructed a novel privacy-preserving scheme which enables mobile users to issue search queries and achieve fine-grained access control over ciphertexts for cloud-based mobile crowdsourcing. In this paper, our research focuses on the study of clustering analysis based on differential privacy protection [23], a new privacy model grounded on data distortion [24].

On one hand, differential privacy, as a promising solution, helps ease the tension of aforementioned data leaks and allows analysts to perform benign aggregation analysis while ensuring that personal privacy is effectively protected; however on the other hand, data authenticity is reduced due to the addition of noise to varying degrees. The goal of our research is to improve the precision of clustering results while protecting the data privacy. The main contributions of this paper are as follows.

1) To ensure the security of data clustering, we introduce the differential privacy and k-means algorithm, and propose a privacy and availability data clustering (PADC) scheme, which improves the selection of the initial center points. The specific method calculates the density of each data point and sorts them, then divides data into segments according to the order of the density. The average of each segment is the center point, which helps to improve the availability of clustering while preserving privacy.
2) We propose an outlier parameter for detecting the outliers of dataset. Removing the outliers in the data clustering process helps to enhance the precision of clustering results.
3) We use the relative distance, rather than the Euclidean distance, to calculate the distance from each point to the center point. This method adds weight to each cluster, helping to make the division of points more accurate.
4) Security analysis indicates that the proposed PADC scheme can resist against the center point attack and the background knowledge attack. The result of performance evaluation shows that the proposed scheme is both effective and efficient.

The rest of this paper is organized as follows. We introduce the related work in Section II, and provide an elaborate description of the proposed PADC scheme's construction in Section III. Our security analysis and design of the experiment evaluation are presented in Sections IV and V, respectively. We conclude this paper in Section VI.

II. Related Work

A. Differential Privacy Mechanism

In recent years there have been many privacy protection methods based on k-anonymous [25] and data division, such as the l-diversity [26] and the t-closeness [27]. The basic notion of these models is to define the attributes of the dataset associated with the attacker's background knowledge as quasi-identifiers, and then generalize and compress the quasi-identifier values of the records so that all records are divided into multiple equivalents class. The records in each equivalence class have the same quasi-identifier value, thereby enabling a record to be hidden in a set of records. Therefore, such models are also known as group-based privacy protection models. Although these methods are capable of protecting the details of data, they requires special attack assumptions and background knowledge. The differential privacy mechanism [28] was proposed by Dwork in 2006. Unlike traditional privacy protection methods, the differential privacy mechanism is a privacy protection technique based on data distortion. The amount of background knowledge from the attacker is irrelevant, and it distorts the sensitive data while keeping some data or data attributes unchanged by adding noises. In order to ensure that the processed data will still retain certain statistical aspects and carry out data mining or other operations, differential privacy defines a rigorous attack model and the risk of privacy disclosure requires a rigorous, quantitative representation and proof. This approach significantly reduces the risk of privacy leaks, while maximally ensuring the availability of data.

As an example, dataset D contains Alice's privacy information. Any random query operation K on D would produce an

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

output $K(D)$. Even if Alice's privacy information is removed from the dataset D, the result of this query operation is still $K(D)$. Therefore, it can be concluded that Alice's privacy information is not exposed. In other words, differential privacy ensures that datasets containing particular data would not affect the results of any query. The specific definition is given as follows.

Definition 1: A randomized function K gives ϵ-differential privacy if datasets D_1 and D_2 differ on at most one element, and all range $\subseteq$ Range(K), we have

$$\Pr\{K(D_1) = \text{range}\} \leq e^{\epsilon} \cdot \Pr\{K(D_2) = \text{range}\}. \tag{1}$$

The definition indicates that the random function K achieves the goal of privacy protection through the random operation to the output result, and the probability of the operation result is independent.

At present, the main purpose is to add noise to the output result, and increase data uncertainty to reduce data authenticity. Researchers proposed the Laplace mechanism [29] which has been generalized to yield a publicly available codebase for writing programs that ensure differential privacy and provide a differentially private interface for accessing data for analysis purpose. Such an interface can be useful even when it is difficult to determine the sensitivity of the desired function or query sequence. It can also be used to run an iterative algorithm, composed of easily analyzed steps, for as much iteration as a given privacy budget permits. In addition, this mechanism applies to numerical data, and achieves the ϵ-differential privacy by adding the random noise that matches the Laplace distribution to the exact query results. The definition of Laplace mechanism is as follows:

$$F(D) = f(D) + \text{Lap}(b) \tag{2}$$

where $b = \Delta f/\epsilon$, Δf denotes the sensitivity of the query function, referring to the maximum distance difference that occurs when the query function f acts on the adjacent dataset, ϵ is the privacy protection parameter, representing the amount of noise. The calculation method is described in Definition 2.

Definition 2: For $f : D \to R^{\text{dim}}$, the input is a dataset and the output is a dim dimensional real vector. For any adjacent datasets D_1 and D_2, the sensitivity of f is defined as

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1. \tag{3}$$

On this basis, Ebadi *et al.* [30] improved more complex personalized privacy mechanism, with which a user can set the protection levels according to the different privacy levels of files. In addition, Geng *et al.* [31] designed a mechanism with additional optimization steps for adding noise.

B. k-Means Algorithm

Clustering algorithms are a category of unsupervised machine learning algorithms and an important part of data mining, serving as the basis of many data mining methods. It can be used to discover the potential value of business decision-making knowledge and rules from a large number of relevant data [32]. The essence of clustering algorithms is to group datasets, not directly comprehended by human, based on their similarity [33] for revealing the realities of data distribution. Clustering algorithms divide the dataset with nodes into multiple classes, each with an initialized central node. The nearest node from each center is divided into different classes according to the calculated distance. Through the iterative method, value of the cluster center is then constantly updated until the convergence function converges. Current clustering algorithms are roughly divided into five categories [32]: 1) the partition method; 2) the hierarchical approach; 3) the density-based approach; 4) the grid-based approach; and 5) the model-based approach. The k-means algorithm is one of the representative partition clustering algorithms. Serving as an important method in data mining technique, it has the advantages of simplicity and fast speed.

C. k-Means Clustering Based on Differential Privacy

By analyzing the k-means algorithm and its privacy leakage, it is not difficult to discover that the key of privacy disclosure is the center point. The clustering center point is obtained by dividing the sum of the data points in a cluster by the number of data points. When the dataset is clustered, the detailed data point information is not needed, with the provision of the approximation of each cluster center point, which is essential in protecting the data privacy while not affecting the precision of the clustering results.

The traditional differential privacy k-means algorithm proposed by Blum *et al.* [34], provides differential privacy protection for the sensitive information in a dataset by adding a small amount of appropriate noise to the center point in the clustering process so that the risk of disclosing the center point satisfies the definition of differential privacy. A large number of simulation experiments however show that the traditional differential privacy k-means algorithm is more sensitive to the selection of the initial center point. Aiming at this problem, Li *et al.* [35] proposed their improved differential privacy (IDP) k-means algorithm, which improves the selection of initial center point. Following their method, the dataset D is divided into k subsets as initial clusters, followed by the calculation of the mean of each cluster and adding noise at the initial center points. It has been proven that the IDP k-means algorithm has significantly improved the availability of clustering results compared with the DP k-means algorithm, under the condition that the noise addition and the privacy budget are constant. In addition, Dwork [36] developed a different setting method of privacy budget and Nissim *et al.* [37] proposed the *PK*-means algorithm, which enables the center of the k-means clustering to satisfy the differential privacy mechanism.

D. Motivation

We have observed the following important issues through the above analysis.

1) The k-means clustering algorithm leads to the uncertainty of clustering results due to the randomness of initial cluster center selection. Additionally, the algorithm is sensitive to outliers: the more the outliers in dataset, the more serious the impact on clustering result.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

TABLE I
NOTATIONS AND DESCRIPTIONS

Notations	Descriptions
D	a dataset
dim	the dimension of the dataset
n	the number of points in dataset
r	the parameter about the outliers
k	the number of the final clusters
c_i	the center point of the i cluster
s_i^2	the variance of the i cluster
w_i	the weight of the i cluster
$dist$	the distance of two points
N	the number of iterations
sum	the sum of points of a cluster
num	the number of points of a cluster
ϵ	the privacy protection budget
i	the count parameter

2) In order to protect the security of data clustering process of *k*-means algorithm, we introduce the differential privacy mechanism [23]. Thanks to it being based on data distortion, the randomness of initial cluster center is increased. Furthermore, the center point of every cluster also deviates from the real one. These factors can lead to more unstable clustering results.

In order to solve the above problems, this paper presents the following solutions, with a comprehensive analysis indicating its security and effectiveness.

1) The detection of outliers. Each dataset will have more or less points away from the dataset, and they will affect the results of clustering. If these outliers can be detected and labeled, the precision of the clustering results will be enhanced.
2) Improved selection of the initial point. The clustering center of existing algorithms are far from the correct one because of noise, making clustering results unavailable. If the selection of the initial center point can be improved, making the initial center point in the close proximity of the correct clustering center point, then the availability of clustering results will be greatly enhanced.
3) The original algorithm uses the Euclidean distance calculation method for computing the distance between two points in the process of clustering. We propose to assign a corresponding weight to each cluster according to the tightness of different clusters in each iteration, and add the weight when calculating the distance to the center of each cluster to achieve more accurate partitioning of points.

III. CONSTRUCTION OF THE PROPOSED SCHEME

Notations and the corresponding descriptions are given in Table I, followed by the proposed system model, the adversary model, and detailed steps of our proposed scheme.

A. System Model

In this section, we introduce and present a three-layer system model [1] as shown in Fig. 1: 1) perception layer; 2) network layer; and 3) application layer for scenarios where electricity consumption data is collected. The main functions of each layer are as follows.

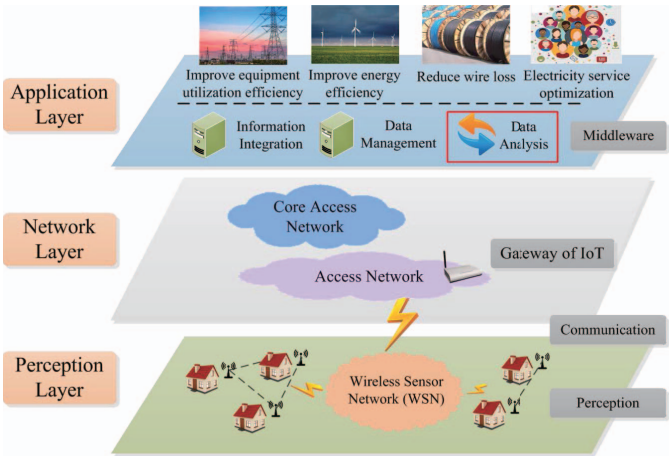


Fig. 1. System model for scenarios collecting electricity consumption data.

1) *Perception layer* implements the collection and processing of household and field electricity consumption data from the physical world. This layer is essentially a WSN, formed by multiple sensor nodes communicating with one another.
2) *Network layer* implements the transmission and control of the electricity consumption data.
3) *Application layer* is responsible for analyzing and processing the electricity consumption data collected by the perception layer to achieve intelligent control, decision-making, and service provision for improving the intelligence of the smart grid in various applications.

The following sections will address the privacy concerns and the precision of *k*-means clustering during data analysis at middleware of application layer.

B. Adversary Model

Upon receiving the user's electricity information at the control center, data analyst clusters the data for better services, such as electrical supply and electrical dispatch decision support and electricity consumption forecast with improved accuracy. Attackers may steal user's privacy information during this process. To simulate such attacks, we propose the following two adversary models.

1) *Attack Based on Center Point:* The attribute values of certain data points may be exposed during the calculation of the distance. For example, suppose that the electricity dataset of a certain cell in a certain month contains three attributes: a) house number; b) family size; and c) electricity consumption. The values of these three attributes form the coordinates of the data point. Table II shows an example of the distances of a data point $a(x, y, z)$ from two center points in two iterations. Assuming that the attacker acquires the information in Table II, the specific value of the data point can be deduced by the simultaneous equations.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXAMPLE CENTER AND DISTANCE INFORMATION

Iterations	Center C_1	Distance d_1	Center C_2	Distance d_2
1	(1,3,322.5)	2.43	(2,4,238)	2.00
2	(3,2,154.3)	3.00	(3,3,123.4)	3.16

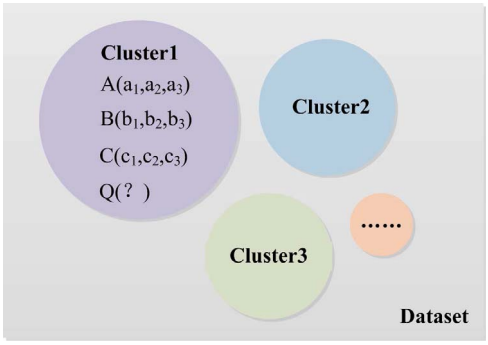


Fig. 2. Example showing cluster information.

More serious privacy exposure can be caused by more interactions with fewer sample attributes. Using information, including house number, number of occupants, and electricity consumption for a month, it can be further inferred whether the house was occupied during a given period of time. The amount of electricity used can even help determine the actual number of residents at the property.

2) *Attack Based on Background Knowledge:* Similar to the first example, the four points represent the electricity consumption of four households in a month. Adversary may be carrying out privacy attacks based on the final cluster center with background knowledge. As shown in Fig. 2, for example, in a data clustering analysis, the four points *A*, *B*, *C*, and *Q* are grouped in the same cluster in an iteration. We assume the adversary has all the background knowledge. It is known that data point *Q* belongs to cluster 1 and all data points, including center point and other than *Q* are known. With this information, an attacker can deduce each attribute value of *Q* according to the clustering mean center point calculation formula.

C. Basic Definition

1) Detection of Outliers: Electricity datasets typically contain unique data points. An example would be minimal electricity consumption of certain households not being at home all year round. Extremely high and low electricity consumption data can affect the analysis of local electricity consumption. The corresponding data points are far apart from the center data point and therefore are referred to as outliers. It is desired to detect and remove these data points during the calculation of the center point in the clustering process for better precision of such calculation and clustering. For this reason, matrix is used to store the distance square of each data point to all other points in the process of finding the initial point. As shown in (4), for each data point, the ratio of the number of data points to the square of the distance represents the point's density value

$$\text{density}(x) = \frac{n}{\sum_{i=1}^{n} \text{dist}^2(x, y_i)}. \tag{4}$$

According to (4), greater density value indicates a more compact proximity around the point. Given outliers may affect the calculation of the center point, density values of all points are sorted in a decreasing manner and outliers can be excluded by removing the ending portion of the sorted queue. Specifically, an outlier parameter r is introduced for such purpose, e.g., if $r = 0.9$, then the number of accepted data points is $(n * 0.9)$, and the rest of the points $(n * 0.1)$ are considered as outliers and therefore labeled. In the subsequent iteration process, while the outliers are still part of the cluster division, they do not participate in the calculation of the center point. To ensure that the appropriate value of r is chosen, experiments are conducted against different datasets as presented in the later sections.

2) Distance Calculation: In each iteration, the similarity of data points inside each cluster may vary. Therefore, a weight is given according to the level of the similarity within the cluster when calculating the distance between a data point and a cluster center. The similarity of the cluster can be measured by the variance of data points. A large variance indicates a low level of similarity of the cluster, while on the contrary, a small variance suggests a high level of similarity. With a high similarity, the Euclidean distance between the data points and the clusters is given a greater weight value, while a low similarity is associated with less weight. Therefore, the weight functions as the reciprocal of variance. Relative distance is used in distance calculation, where variance is employed for examining the variation of all data points in the cluster, and thus is susceptible to outliers. In order to reduce or eliminate the influence of outliers, a number of them are removed from the dataset when calculating the variance. Specifically, only 90% of the data points with smaller distance to the center of the cluster are retained for calculating the variance, which is expected to be more accurate and better reflect the actual dispersion degree of the cluster. This can be formulated as

$$s_i^2 = \frac{\sum_{x \in c_i} \sum_{i-1}^{n \times 0.9} (x - c_i)^2}{n_{c_i}}. \tag{5}$$

Based on the variance, the weight of this cluster can be obtained, as shown in

$$w_i = \frac{1}{s_i^2}. \tag{6}$$

Subsequently, the aforementioned relative distance calculation can be formulated as

$$\text{dist}^2(x, c_i) = w_i \cdot \sum_{i=1}^{d} (x_i - c_i)^2. \tag{7}$$

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Choose the Initial Center Points

input: the original dataset D, the number of clusters k, the outlier parameter r
output: k center points
for $i \leftarrow 0$ to *data.length* **do**
  compute the *density*[i] of *data*[i];
end for
sort(*density*)from large to small;
for $j \leftarrow data.length * (1 - r)$ to *data.length* **do**
  outlier[j] = *data*[j];
end for
int *number* = $data.length * r/k$;
for $i \leftarrow 0$ to k **do**
  for $j \leftarrow i * number$ to $(i + 1) * number$ **do**
    centers[i] + = *data*[j];
  end for
  centers[i] = *centers*[i]/s;
end for

D. Scheme Description

The proposed PADC scheme consists of the following two principle phases.

Phase I: Compute the density of each data point for identifying outliers, then divide data into segments for finding the center point according to the order of the density. Algorithm 1 shows the pseudo-code of Phase I.

1) Traverse all points, and calculate the square value of the distance from each point to other points.
2) Calculate the density value of each point using (4).
3) Sort all density values in the order of large to small.
4) Based on the outlier parameter r, label and process the $n * (1 - r)$ data points at the end of the sorted queue as outliers.
5) Based on the value of k, the sorted data points, excluding the outliers, are divided into k clusters; the center of each cluster is made as the initial center point.

Phase II: Use the selected initial center points from Phase I for clustering the dataset. The pseudo-code is presented in Algorithm 2.

1) Calculate the Euclidean distance from each point in the dataset D to each center point and divide it into the nearest center point according to the nearest Euclidean distance, resulting in k clusters as the initial classification.
2) After excluding the prelabeled outliers, the total number of remaining data points is $n*r$. Calculate the number of data points *sum* and the total number of data points *num* in each cluster, then add noise Lap(b). The center point of the cluster is then updated as $(\text{sum}+\text{Lap}(b))/(\text{num}+\text{Lap}(b))$.
3) According to (5) and (6), calculate the weight of each cluster using (7) of the distance calculation to retraverse all points for clustering the division.
4) Repeat the previous two steps until the convergence function is converged.

Algorithm 2 Clustering

input: the original dataset D, k initial center points, the number of clusters k, the sensitivity of query function Δf, ϵ value
the outlier parameter r
output: clustering result
compute the *weight*[i] of *clusters*[i];
while the convergence function is converges **do**
  for $i \leftarrow 0$ to k **do**
    compute the *weight*[i] of *clusters*[i];
  end for
  for $x \leftarrow 0$ to *data.length* **do**
    for $y \leftarrow 0$ to k **do**
      $a[y] = weight[x] * dist(S[x], initial[y])$;
    end for
    find the minimum value of a;
    categorize S[x] to the nearest center point;
  end for
  for $i \leftarrow 0$ to k **do**
    sum = the sum of data points of the i cluster;
    num = the number of data points of the i cluster;
    $sum' = sum + Lap(sensitivity/\epsilon)$;
    $num' = num + Lap(sensitivity/\epsilon)$;
    $centerpoint = sum'/num'$;
  end for
end while

E. Privacy Protection Parameter Setting

According to the PADC scheme, the center point of the cluster is $(\text{sum} + \text{Lap}(b))/(\text{num} + \text{Lap}(b))$. The sensitivity of the denominator is 1, and the sensitivity of the numerator is determined by the dimension of the dataset. Adding or deleting a point in the dim dimension dataset, the sensitivity of each dimension is 1, and the sensitivity of the numerator is dim, then the sensitivity of the entire query sequence is $(\text{dim} + 1)$.

Different datasets have different iterations in the clustering algorithm, and two methods of setting the privacy budget ϵ proposed by Dwork [36] are as follows.

1) If the number of iterations N is known, the privacy budget for each consumption is ϵ/N, according to Definition 2, the size of each noise added is $\text{Lap}((\text{dim} + 1) * \epsilon/N)$.
2) If the number of iterations N is unknown, the value of the parameter ϵ can be adjusted continuously during the iteration.

Based on our great amount of experience, the effect of preiteration on clustering results is greater than the post-iteration. Consequently, our experiment chooses to increase the parameters ϵ gradually in the clustering process. The first consumption is $\epsilon/2$ with noise $\text{Lap}(2(\text{dim} + 1)/\epsilon)$, and each iteration consumption of the budget is half of the latest iteration until the end of the iteration.

IV. Security Analysis

If the random function in the scheme can provide the noise to satisfy the Laplace distribution, then it can also provide the

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

differential privacy protection to the query result. The scheme proposed in this paper provides ϵ-differential privacy protection to the clustering result through satisfying the Laplace distribution to the center points in the process of clustering by adding appropriate noise.

Following the differential privacy definition, use PADC(D) to represent the result of PADC k-means scheme clustering for dataset D, and use d to represent the center point of dataset D. Then we have Lemma 1.

Lemma 1: The PADC function gives ε-differential privacy if two datasets D_1 and D_2 differ on at most one element and all range $\subseteq$ Range(PADC). We have

$$\Pr\{\text{PADC}(D_1) = \text{range}\} \leq e^{\varepsilon} \cdot \Pr\{\text{PADC}(D_2) = \text{range}\}. \quad (8)$$

Proof: The Laplace function provides differential privacy protection for the clustering center points. Therefore, for the center points c_1 and c_2 of the two datasets D_1 and D_2, the risk of disclosure satisfies

$$\Pr\{\text{Lap}(c_1)\} \leq e^{\varepsilon} \cdot \Pr\{\text{Lap}(c_2)\}. \quad (9)$$

In the original K-means algorithm, it is assumed that the attacker obtains all the information except the target information in the cluster and the center points where noise was added. Consequently the attacker can obtain the accurate sensitive attributes of the data sample by calculating the distance between data samples to the center point. Therefore, the accuracy of the center point directly affects the precision of the target information, and the probability of obtain the accurate information about the dataset is equal to the exposure risk of the accurate center point. Then we can have (10) and (11), which coincide with (8) for differential privacy. In other words, the PADC scheme can provide ϵ-differential privacy protection for the dataset

$$\Pr\{\text{PADC}(D_1)\} = \Pr\{\text{Lap}(c_1)\} \quad (10)$$

$$\Pr\{\text{PADC}(D_2)\} = \Pr\{\text{Lap}(c_2)\}. \quad (11)$$

In defending against the two types of attacks aforementioned in the adversary model, random noise is added to the center point through the Laplace mechanism in each iteration. Consequently, the adversary will not be able to deduce each attribute value of the data point based on the clustering mean center point calculation formula, even if the adversary has acquired the distances between a data point and two center points in two iterations or the background knowledge about the data point.

Therefore, in the real world scenario, the adversary is incapable of learning the details of electricity consumption of a specific household and thus will not be able to draw conclusions on the occupancy of the property, which helps protect the security and privacy of residents. ■

V. Performance Evaluation

We conducted a set of simulation experiments with Java on a Windows 7 test computer with the following configurations: Intel Core i5-4590 3.3 GHz CPU, 8 GB RAM, and 1 TB hard Disk. In order to demonstrate the effectiveness of the proposed PADC scheme, we employ four common datasets running with the PADC k-means, DP k-means [34], and IDP k-means [35] algorithms based on differential privacy preservation. The results are compared and evaluated in this section.

TABLE III
Experimental Data Information

Dataset	Dims	Type	Tuples	Clusters
Iris	4	Real	150	3
Wine	13	Real	178	3
Climate	18	Real	540	2
Gamma	11	Real	19020	2

A. Complexity Analysis

In the aforementioned scheme, when the perception layer collects electricity consumption data from different houses and transmits it to the network layer and application layer, the infrastructure part of the application layer performs cluster analysis on the data. In this process, differential privacy protection mechanism is used to protect information privacy. The difference from the k-means algorithm is introduced by adding appropriate amount of noise to the center point during the iteration process, so that the clustering process will be biased, and the number of iterations N may increase. The data may have a clustering structure, the number of iterations until convergence is typically small, and results only improve slightly after the first dozen of iterations. Regarding computational complexity, if the final clusters k and the dimension d are fixed values, the problem can be solved in time $O(n{*}d{*}k{*}N)$, where n is the number of entities to be clustered, N is the number of iterations, and spatial complexity is $O(n*d)$. Furthermore, in general, d, k, and N can all be considered constants, so the time and space complexity can be simplified as $O(n)$, where computational complexity is then linear to the number of data points in the dataset.

B. Dataset

In order to reflect the advantages of our scheme, we employ different types of datasets in our simulation. The datasets used in the scheme are from the UCI Knowledge Discovery Archive database (http://archive.ics.uci.edu/ml/datasets.html). The basic information of the data is shown in Table III.

C. Evaluation Indicator

Since the datasets used are already available and accessible in clustering results, only external evaluation indicator is needed for evaluating the effectiveness of clustering results. F-measure, a combination of precision and recall, is used here to measure the clustering usability. The Iris dataset classification in the UCI database is taken as an example to accurately describe the evaluation index, as shown in Table IV, the number of data points in different cases is represented using a variable. Setosa, Versicolor, and Virginia are the names of the three categorizers in the Iris dataset. The horizontal axis represents the true result of the clustering algorithm, and the vertical axis represents the correct classification of the dataset.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
NOTATIONS AND DESCRIPTIONS OF THE IRIS DATASET CLASSIFICATION

	Setosa	Versicolor	Virginia
Setosa	A	B	C
Versicolor	D	E	F
Virginia	G	H	J

From Table III, it is not difficult to obtain precision and recall of different classification. Precision and recall rates are two metrics that are widely used in the field of information retrieval and statistical classification. In general, precision is the ratio, in the algorithm result, of the number of data points that are correctly clustered to the total number of data points. Recall is the ratio of the number of data points correctly clustered in the result of the algorithm to the total number of data points in the dataset. For Setosa in the above example, its precision and recall are

$$\text{Precision(Setosa)} = \frac{A}{A+D+G} \tag{12}$$

$$\text{Recall(Setosa)} = \frac{A}{A+B+C}. \tag{13}$$

Consequently, the F-Measure value of the cluster can be obtained from

$$F(\text{Setosa}) = \frac{2 \times \text{Precision(Setosa)} \times \text{Recall(Setosa)}}{\text{Precision(Setosa)} + \text{Recall(Setosa)}}. \tag{14}$$

Given that a dataset is usually divided into multiple clusters, for the whole clustering result of dataset D, the F-measure value is then formulized in (15), where $|c_i|$ represents the number of data points of the ith cluster

$$F = \frac{\sum_{i=1}^{k}(|c_i| \times F(i))}{\sum_{i=1}^{k}|c_i|}. \tag{15}$$

Our research implements a differential privacy k-means clustering operation on the dataset under different privacy budgets. The F-measure value of the clustering result is then calculated. When the results of the two clusters are the same, the result of F-measure takes the maximum value of 1. The larger the result of F-measure, the greater the similarity of the two clustering results, and the smaller the effect of the noise added by the differential privacy protection mechanism on clustering usability. In the following section, we compare the experimental results with the clustering results provided by the UCI dataset to determine the usability of the clustering results.

D. Experimental Results

1) Outlier Parameter r: In our scheme, the value of parameter r is an important factor affecting the stability of clustering algorithm. With an appropriate value of r, the outlier can be labeled, and thus ensuring the stability of the clustering algorithm. Therefore in order to stabilize the performance of the experiment, an appropriate value of r has to be determined. Four datasets were clustered without adding a differential privacy noise mechanism. Given that outliers only count a relatively small portion, the range of r is set to [0.7, 1] with an interval of 0.02. The experimental results are shown in Fig. 3.

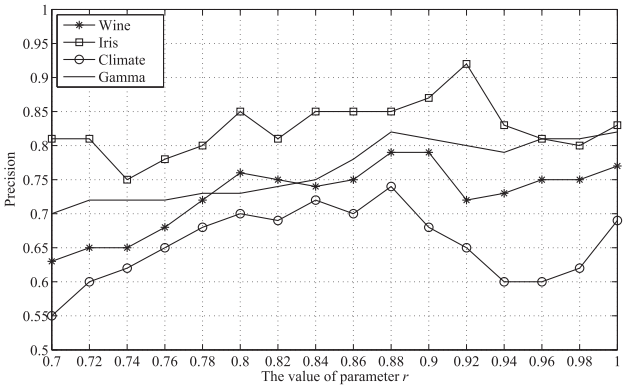


Fig. 3. Precision with various r values.

The abscissa represents the value of r, and the ordinate represents the precision of the clustering algorithm over different datasets. Fig. 3 shows that the different values of r correspond to the different precision of the clustering. This is because the value of parameter r leads to different population portions of outliers, resulting the diverse effect on the clustering results. Another finding is that there is no gradual increase or decrease in the linear relationship.

In addition to the above discoveries, we observed that certain normal points are categorized as outliers due to the corresponding low density values. For this reason, an appropriate value of r needs to be determined to prevent such fault. Fig. 3 also shows that the best precision of clustering results is achieved when the value of r falls in the range of [0.88, 0.92], which corresponds to 8%–12% points being outliers.

Taking the dataset Wine as an example, the peak of precision is found with the value of r between 0.88 and 0.90, similarly for Iris with $r = 0.92$. For both Climate and Gamma, the r value of 0.88 gives the best precision. Based on the above experimental results, a r value of 0.90, median of the interval, is chosen to apply to most datasets. In other words, each dataset will have $n * 0.1$ points labeled as outliers.

2) Clustering Performance Analysis: In the following cluster analysis experiment, PADC k-means, DP k-means [34], and IDP k-means [35] are used against each dataset. Errors may be introduced due to the randomness of noise in the experiment. In order to minimize possible errors, the value of ϵ is gradually increased from 0.1 in this experiment for each dataset. Each ϵ value runs 20 times to obtain an average of F-measure. The experimental results are presented in Figs. 4–7.

The final F-measure of PADC k-means algorithm has increased significantly with the same ϵ value. The improvement of the clustering availability indicates the superior performance of the PADC k-means scheme. Additionally, as ϵ increases, the added noise decreases and the data availability of PADC k-means scheme gradually approaches to that of the original k-means algorithm.

IDP k-means algorithm divides the dataset into several subsets as initial clusters, within each of which the data points may not be in the same cluster. For this reason, it can be seen from the figures that the availability of clustering results is not optimal with small ϵ value and large amount of noise, even though it still outperforms DP k-means algorithm.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

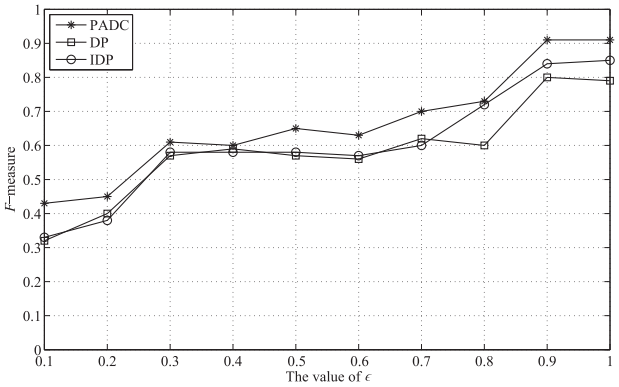


Fig. 4. F-measure with various ϵ values in Wine dataset.

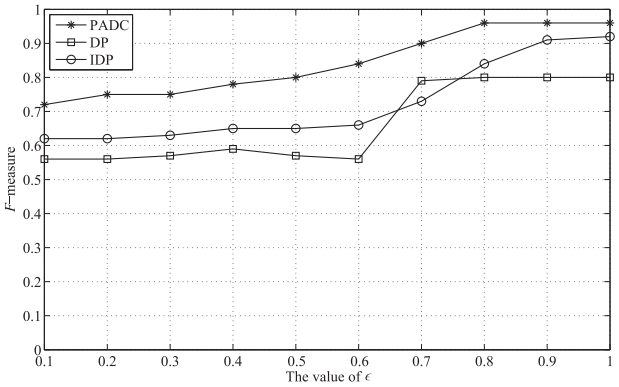


Fig. 5. F-measure with various ϵ values in Iris dataset.

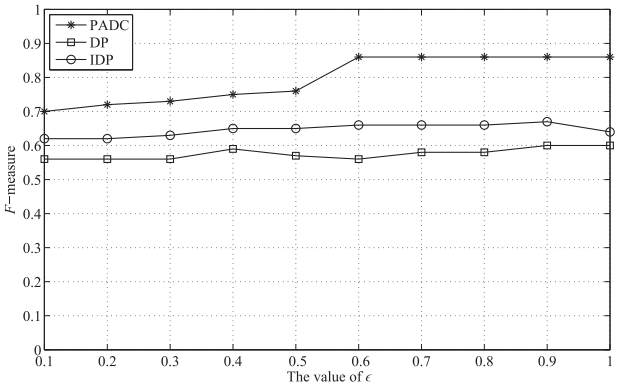


Fig. 6. F-measure with various ϵ values in Climate dataset.

Our proposed scheme not only improves the selection of the initial points but also detects the outliers and adds weight of cluster to the distance calculation. These factors greatly enhance the stability of initial classification, subsequently resulting improved accuracy of clustering centers and reduced deviation of clustering results.

In summary, our experimental results show that the proposed PADC k-means clustering algorithm is superior to both the IDP k-means and DP k-means algorithms in terms of clustering effectiveness and efficiency at the same level of privacy.

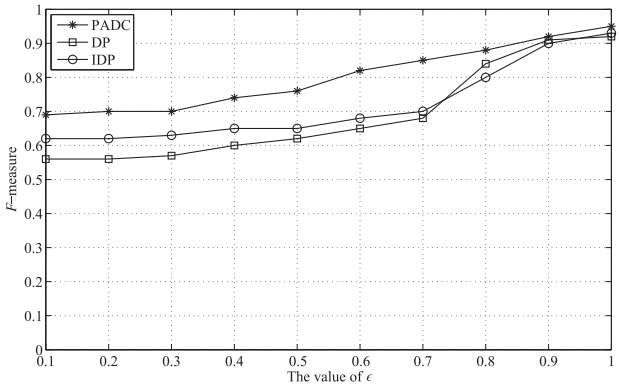


Fig. 7. F-measure with various ϵ values in Gamma dataset.

VI. Conclusion

In an effort of providing better electricity services, providers utilize WSNs for collecting electricity consumption data for data clustering analysis. Such privacy data may be subject to disclosure in the process of data clustering. While existing solutions using differential privacy mechanism may protect data privacy, they suffer in providing availability of data clustering results. In this paper, we proposed the PADC scheme for ensuring the available of data clustering results under the premise of security through improving the selection of the initial cluster centers and finding the outliers via calculating the density of every data point. In our PADC scheme, weights are added to each cluster according to the cluster density in each iteration, which assists in calculating the relative distance among data points for more accurate division results in the same iteration. Experiment results show that our proposed scheme improves the availability of clustering results when compared to other differential privacy k-means algorithms with the same privacy level. Future work includes optimizing our proposed scheme for enhanced precision of data clustering results.

References

[1] J. Lin *et al.*, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[2] D. Wu, Q. Liu, H. Wang, D. Wu, and R. Wang, "Socially aware energy-efficient mobile edge collaboration for video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2197–2209, Oct. 2017.

[3] R. Wang, J. Yan, D. Wu, H. Wang, and Q. Yang, "Knowledge-centric edge computing based on virtualized D2D communication systems," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 32–38, May 2018.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[5] Q. Jiang, S. Zeadally, J. Ma, and D. He, "Lightweight three-factor authentication and key agreement protocol for Internet-integrated wireless sensor networks," *IEEE Access*, vol. 5, pp. 3376–3392, 2017.

[6] D. Wu, J. Yan, H. Wang, D. Wu, and R. Wang, "Social attribute aware incentive mechanism for device-to-device video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1908–1920, Aug. 2017.

[7] Z. Zhang, P. Zhang, D. Liu, and S. Sun, "SRSM-based adaptive relay selection for D2D communications," *IEEE Internet Things J.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8026120/, doi: 10.1109/JIOT.2017.2749443.

[8] J. Guo, J. Ma, X. Li, T. Zhang, and Z. Liu, "A situational awareness trust evolution model for mobile devices in D2D communication," *IEEE Access*, vol. 6, pp. 4375–4386, 2018.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

[9] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between Internet of Things and social networks: Review and research challenges," *IEEE Internet Things J.*, vol. 1, no. 3, pp. 206–215, Jun. 2014.

[10] I. V. Cadez, P. Smyth, and H. Mannila, "Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction," in *Proc. 7th. Int. Conf. Knowl. Disc. Data (SIGKDD)*, San Francisco, CA, USA, Aug. 2001, pp. 37–46.

[11] L. Xu, R. Collier, and G. M. P. O'Hare, "A survey of clustering techniques in WSNs and consideration of the challenges of applying such to 5G IoT scenarios," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1229–1249, Oct. 2017.

[12] X. Liu, R. Deng, K.-K. R. Choo, Y. Yang, and H. Pang, "Privacy-preserving outsourced calculation toolkit in the cloud," *IEEE Trans. Depend. Secure Comput.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8318625/, doi: 10.1109/TDSC.2018.2816656.

[13] Y. Yang, X. Liu, R. H. Deng, and Y. Li, "Lightweight sharable and traceable secure mobile health system," *IEEE Trans. Depend. Secure Comput.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7987011/, doi: 10.1109/TDSC.2017.2729556.

[14] J. Xiong *et al.*, "RSE-PoW: A role symmetric encryption PoW scheme with authorized deduplication for multimedia data," *Mobile Netw. Appl.*, vol. 23, no. 3, pp. 650–663, 2018.

[15] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Proc. 20th. Int. Conf. Cryptograph. (CRYPTO)*, Santa Barbara, CA, USA, Aug. 2000, pp. 36–54.

[16] L. Chen, J. Ji, and Z. Zhang, *Wireless Network Security: Theories and Applications*. New York, NY, USA: Springer, Sep. 2013.

[17] J. Vaidya and C. Clifton, "Privacy-preserving *k*-means clustering over vertically partitioned data," in *Proc. 9th. Int. ACM Conf. Symp. Knowl. Disc. Data Min. (SIGKDD)*, Washington, DC, USA, Aug. 2003, pp. 206–215.

[18] S. Jha, L. Kruger, and P. D. McDaniel, "Privacy preserving clustering," in *Proc. 10th. Int. Conf. Symp. Res. Comput. Security (ESORICS)*, Milan, Italy, Sep. 2005, pp. 397–417.

[19] A. Inan *et al.*, "Privacy preserving clustering on horizontally partitioned data," *Data Knowl. Eng.*, vol. 63, no. 3, pp. 646–666, 2007.

[20] J. Regenold, K. Wang, G. Smith, Q. Liu, and L. Chen, "Enhancing enterprise security through cost-effective and highly customizable network monitoring," in *Proc. 10th EAI. Int. Conf. Mobile Multimedia Commun. (MOBIMEDIA)*, Chongqing, China, Jul. 2017, pp. 133–142.

[21] D. Wu, S. Si, S. Wu, and R. Wang, "Dynamic trust relationships aware data privacy protection in mobile crowd-sensing," *IEEE Internet Things J.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/8089752/, doi: 10.1109/JIOT.2017.2768073.

[22] Y. Miao *et al.*, "Practical attribute-based multi-keyword search scheme in mobile crowdsourcing," *IEEE Internet Things J.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8125665/, doi: 10.1109/JIOT.2017.2779124.

[23] J. Ren, J. Xiong, Z. Yao, R. Ma, and M. Lin, "DPLK-means: A novel differential privacy K-means mechanism," in *Proc. Int. Conf. DSC*, Shenzhen, China, Jun. 2017, pp. 133–139.

[24] C. Dwork, "Differential privacy: A survey of results," in *Proc. 5th. Int. Conf. Theory Appl. Models Comput. (TAMC)*, Xi'an, China, Apr. 2008, pp. 1–19.

[25] L. Sweeney, "*k*-anonymity: A model for protecting privacy," *Int. J. Uncertainty Fuzziness Knowl. Based Syst.*, vol. 10, no. 5, pp. 1–14, 2002.

[26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "*L*-diversity: Privacy beyond *k*-anonymity," in *Proc. Int. Conf. Data. Eng. (ICDE)*, Atlanta, GA, USA, Apr. 2006, p. 24.

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. 23rd. Int. Conf. Data. Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[28] C. Dwork, "Differential privacy," in *Proc. 23rd Int. Colloquium Automata Lang. Program. (ICALP)*, Venice, Italy, Jul. 2006, pp. 1–12.

[29] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Int. Conf. Theory Cryptography (TCC)*, New York, NY, USA, Mar. 2006, pp. 265–284.

[30] H. Ebadi, D. Sands, and G. Schneider, "Differential privacy: Now it's getting personal," in *Proc. 42nd Int. ACM Conf. SIGPLAN-SIGACT Principles Program. Lang. (POPL)*, Mumbai, India, Jan. 2015, pp. 69–81.

[31] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1176–1184, Oct. 2014.

[32] J. Sun, J. Liu, and L. Zhao, "Clustering algorithms research," *J. Softw.*, vol. 19, no. 1, pp. 48–61, 2008.

[33] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," *Technometrics*, vol. 32, no. 2, pp. 227–229, 1988.

[34] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SuLQ framework," in *Proc. 24th Int. ACM Conf. SIGACT-SIGMOD-SIGART Symp. Principles Database Syst. (SPDS)*, Baltimore, MD, USA, Jun. 2005, pp. 128–138.

[35] Y. Li, Z. Hao, and W. Wen, "Research on differential privacy preserving k-means clustering," *Comput. Sci.*, vol. 59, no. 1, pp. 1–34, 2013.

[36] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, 2011.

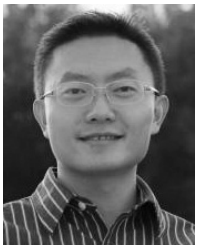
[37] K. Nissim, S. Raskhodnikova, and A. D. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proc. 39th Int. ACM Conf. Symp. Theory. Comput. (STOC)*, San Diego, CA, USA, Jun. 2007, pp. 75–84.

Jinbo Xiong (GS'13–M'14) received the M.S. degree in communication and information systems from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2006 and the Ph.D. degree in computer system architecture from Xidian University, Xi'an, China, in 2013.

He is currently an Associate Professor with the College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China. He has authored or co-authored over 40 publications, 8 patents, and 1 monograph. His current research interests include cloud data security, privacy protection, and mobile Internet security.

Jun Ren received the B.S. degree in software engineering from Fujian Normal University, Fuzhou, China, in 2015, where she is currently pursuing the M.S. degree in software engineering.

Her current research interests include differential privacy and cloud computing security.



Lei Chen (S'04–GS'05–M'06) received the B.Eng. degree in computer science and applications from the Nanjing University of Technology, Nanjing, China, in 2000, and the Ph.D. degree in computer science and software engineering from Auburn University, Auburn, AL, USA, in 2007.

He is currently an Associate Professor, the Interim Department Chair, and the Graduate Program Director with the Department of Information Technology, Georgia Southern University, Statesboro, GA, USA. He has authored or co-authored over 100 peer-reviewed scholarly works. His current research interests include network, information, cloud, and big data security, digital forensics, and mobile, handheld, and wireless security.

Zhiqiang Yao received the Ph.D. degree in computer system architecture from Xidian University, Xi'an, China, in 2014.

He is currently a Professor with Fujian Normal University, Fuzhou, China. He has authored or co-authored over 80 research papers and holds 10 patents. His current research interests include security in cloud computing and multimedia security.

Dr. Yao is a Professional Member of the ACM and a Senior Member of the CCF.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.

Mingwei Lin received the B.S. degree in software engineering and Ph.D. degree in computer science and technology from Chongqing University, Chongqing, China, in 2009 and 2014, respectively.

He is currently an Associate Professor with the College of Mathematics and Informatics, Fujian Normal University, Fuzhou, China. He has authored or co-authored over 20 research papers. His current research interests include storage systems and embedded systems.

Dr. Lin was a recipient of the CSC-IBM Chinese Excellent Student Scholarship in 2012.

Dapeng Wu (M'15–SM'16) is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. He has authored over 100 publications and 2 books. He is the inventor and co-inventor of 28 patents and patent applications. His current research interests include social computing, wireless networks, and big data.

Prof. Wu serves as the TPC Chair of the 10th Mobimedia and Program Committee member of numerous international conferences and workshops. He served or is serving as an Editor and/or a Guest Editor for several technical journals such as *Digital Communications and Networks* (Elsevier) and *ACM/Springer Mobile Network and Applications*.

Ben Niu received the B.S. degree in information security and M.S. and Ph.D. degrees in cryptography from Xidian University, Xi'an, China, in 2006, 2010, and 2014, respectively.

He is currently a Research Assistant with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He was a Visiting Scholar with Pennsylvania State University, State College, PA, USA, from 2011 to 2013. His current research interests include wireless network security and privacy computing.

Authorized licensed use limited to: Carleton University. Downloaded on April 12,2021 at 15:35:25 UTC from IEEE Xplore. Restrictions apply.