An efficient data aggregation scheme with local differential privacy in smart grid

Na Gai [a], Kaiping Xue [a,b,*], Bin Zhu [a], Jiayu Yang [a], Jianqing Liu [c], Debiao He [d]

[a] School of Cyber Science and Technology, University of Science and Technology of China, Hefei, Anhui, 230027, China
[b] Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui, 230027, China
[c] Department of Electrical and Computer Engineering, University of Alabama in Huntsville, Huntsville, AL, 35899, USA
[d] Key Laboratory of Aerospace Information Security and Trusted Computing of Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, Beihu, 430072, China

A B S T R A C T

By integrating the traditional power grid with information and communication technology, smart grid achieves dependable, efficient, and flexible grid data processing. The smart meters deployed on the user side of the smart grid collect the users' power usage data on a regular basis and upload it to the control center to complete the smart grid data acquisition. The control center can evaluate the supply and demand of the power grid through aggregated data from users and then dynamically adjust the power supply and price, etc. However, since the grid data collected from users may disclose the user's electricity usage habits and daily activities, privacy concern has become a critical issue in smart grid data aggregation. Most of the existing privacy-preserving data collection schemes for smart grid adopt homomorphic encryption or randomization techniques which are either impractical because of the high computation overhead or unrealistic for requiring a trusted third party.

In this paper, we propose a privacy-preserving smart grid data aggregation scheme satisfying Local Differential Privacy (LDP) based on randomized responses. Our scheme can achieve an efficient and practical estimation of power supply and demand statistics while preserving any individual participant's privacy. Utility analysis shows that our scheme can estimate the supply and demand of the smart grid. Our approach is also efficient in terms of computing and communication overhead, according to the results of the performance investigation.

1. Introduction

The smart grid is the next-generation power grid that combines modern Information and Communication Technology (ICT) to provide more intelligent services, such as end-to-end connectivity and real-time data management [1]. Smart meters installed in each home communicate electricity consumption data to the control center regularly in a smart grid. The control center collects data from smart meters, analyzes it statistically, and supervises the smart grid's electricity generation, transmission, and distribution [2–4]. Through the smart grid, the control center can estimate the power consumption of the grid and formulate a dynamic pricing strategy.

Despite the promise of the smart grid, the reported data from smart meters always contain private information from consumers. It is possible to analyze a user's electricity usage pattern using these data, posing a serious threat to his or her privacy [5–7]. For example, a user's daily routine can be easily inferred from the electricity usage pattern, and adversaries can analyze whether he/she is at home or not.

Many privacy-preserving data aggregation schemes in the smart grid have been proposed. For example, homomorphic encryption is commonly employed in data aggregation to protect data privacy [3, 8–12]. Homomorphic encryption allows entities to convert plaintext operations into matching ciphertext operations. The smart meters encrypt the data and send the ciphertext to the aggregator using homomorphic encryption, particularly semi-homomorphic encryption such as the Paillier cryptosystem, which permits additional operations on the ciphertext. Then the aggregator integrates the ciphertext gathered from smart meters and decrypts the aggregation result. Homomorphic encryption-based schemes make it possible to preserve a single user's data privacy. However, a common issue is that homomorphic encryption

places a significant computational burden on smart meters, which typically lack appropriate processing power. Moreover, given the fact that smart meters submit data periodically and frequently, homomorphic encryption is not a practical solution for privacy preservation.

Another technique for preserving data privacy in smart grid is to use data masking [13–16]. In these schemes, submitted power data is protected by masking values. Usually, there exists an entity distributing a series of masking values to the smart meters and aggregator. Each smart meter obfuscates the data with the masking value. Then the aggregator can obtain the true result by eliminating the masking values. An exemplary realization of this technique is through Differential Privacy (DP) [17,18]. DP[19] is a mathematically rigorous framework for privacy protection that has been used in a variety of large-scale data aggregation and processing applications that require privacy protection. Random noise, such as Laplacian noise, is commonly used to mask data presented in DP methods. The need for a trusted third party (i.e., a curator) to distribute the noise value is a common difficulty in these schemes. However, this assumption is not always feasible. Although some solutions do not require the use of a trusted third party, distributed approaches usually result in increased user communication overhead.

Local Differential Privacy (LDP) [20–25] has recently gotten a lot of attention in academia and industry. LDP's main idea is that users perform local random perturbations on their data. As a result, without relying on a trusted third party, local differential privacy allows for privacy-preserving data collection and aggregation. Furthermore, as compared to techniques that use homomorphic cryptosystems, the computation overhead is substantially lower. A typical example of LDP implementation is the RAPPOR [23] developed by Google. RAPPOR enables the Google browser to collect statistical information from end-users while providing strong privacy protection. However, most LDP schemes focus on frequency estimation and distribution estimation and mainly deal with discrete category data.

We propose a practical and efficient privacy-preserving data aggregation technique for demand estimate, taking into account the need for data privacy protection and the limited computation capabilities of smart meters in the smart grid (in numerical values). Data privacy can be preserved without the use of a trusted third party by including LDP in data aggregation. When compared to approaches based on encryption algorithms, the processing overhead for each smart meter is acceptable. Furthermore, our scheme has built-in support for users' dynamic joining and quitting, which comes at a low additional cost. At the same time, we investigate the non-general scenario, based on the previous conference version, through which our scheme may meet the more typical situation of smart grid data aggregation supply and demand estimation. The major contributions of our scheme are as follows:

● We propose an LDP-based lightweight privacy-preserving data aggregation scheme in the smart grid, in which smart meters can perturb their generated data by randomized response locally without involving a trusted third party, and the scheme can effectively support users dynamically joining and exiting without involving much extra overhead.

● To assure the usability of aggregated data, we develop a simple but effective data discretization algorithm based on the conditional probability that may reduce the difference between the aggregation results and the real data aggregation results. In such a way, our scheme largely increases the statistical accuracy of data aggregation results. Furthermore, based on the suggested scheme, we analyze more unique cases in smart grid data aggregation and present a grouping approach for large-scale data aggregation, allowing our scheme to handle the majority of smart grid scenarios.

● We implement our proposed scheme on a typical processor and perform performance and security analysis, which shows that the proposed solution has less computation and communication overhead while ensuring utility and privacy protection.

This paper inherits the basic idea of our conference paper [26], which received the best track paper in MSN 2020. They differ mainly in the following aspects: We further consider the scenario of industrial power consumption data aggregation and analysis, where the data range is larger. To ensure the utility of aggregation results, we propose the grouping approach to improve the usefulness of aggregation results, in which the complete data range is divided into different groups, to ensure the usability of aggregation results. When the users perturb their data locally, they first determine which group their data belong to and then perturb the data in a specific group according to the protocol. Meanwhile, we implement our extended scheme on a typical processor and analyze the utility of the extended scheme. We conduct more analysis to further compare the data aggregation error between the un-grouped and the grouped cases.

The rest of this paper is organized as follows. The related work of privacy-preserving data aggregation schemes in smart grid and LDP is given in Section 2. Then we describe the problem model, including the network model and security assumption of our work in Section 3. Preliminaries related to our scheme are given in Section 4. And details of our LDP-based data aggregation scheme are shown in Section 5. In Section 6, we further propose a more universal privacy preserving data aggregation scheme considering more special scenarios in smart grid. Privacy and utility analysis are given in Section 7. Then we give the performance and security analysis in Section 8. Finally, we conclude our work in Section 9.

2. Related work

2.1. Privacy-preserving data aggregation in the smart grid

Data aggregation is a basic service in the smart grid, and privacy protection is one of its primary considerations. To this end, many privacy-preserving data aggregation schemes have been proposed [27, 28]. Homomorphic encryption is one of the most popular methods adopted in privacy-preserving data aggregation schemes, which allows computation on the ciphertext. In Ref. [11], Paillier cryptosystem [29] is introduced to construct a privacy-preserving aggregation scheme for a secure smart grid. In what follows, many other Refs. [9,30] based on Paillier cryptosystem have proposed to protect data privacy under various conditions in smart grid. Some schemes are based on other public-key homomorphic encryption schemes, such as Boneh-Goh-Nissim (BGN) homomorphic encryption algorithm [31,32] and lattice algorithm [2]. Despite its effectiveness in preserving privacy, a practical concern is that the public key-based homomorphic encryption brings too much computation overhead to the smart meters. The smart meter installed in a user's house has limited computing resources, which is costly to conduct encryption functions. Moreover, data acquisition in the smart grid is quite frequent, which means that the smart meter must run encryptions frequently. Therefore, it is impractical to utilize homomorphic encryption to protect data privacy in the smart grid scenario.

Researchers also proposed some schemes based on data masking [13, 15,18,33,34]. In these schemes, the data submitted by users are masked by a masking value, thus the other entities cannot access the real value without knowing the masking value. In Refs. [15,34], schemes satisfying differential privacy were proposed. These schemes reduce the computation overhead on smart meters while achieving privacy protection. However, in some of these masking schemes such as [15], a trusted third party is needed for generating and distributing the masking value. This brings in a new problem that it is hard to find such a trusted party in the real world. There are also some distributed schemes [14] that do not depend on a trusted third party. The masking value is generated by negotiation between users, but it increases the communication cost between users. Besides, existing DP-based schemes are inefficient when it comes to the changing set of users. That is to say, when a user joins or leaves the system, new values should be generated and distributed, increasing the communication overhead.

2.2. Local differential privacy

Local differential privacy [20,35] has been proposed to provide privacy protection for distributed scenarios where users perturb their data locally and upload it without any trusted third party. Different from a centralized differential privacy framework, in which the data are perturbed at the aggregator side. Users perform randomized perturbations on their data locally without noise dispersed from other entities in the local differential privacy framework, ensuring that the aggregator has no access to the user's original data. At present, most of the schemes satisfying LDP are realized by the Randomized Response (RR) [36,37]. Other schemes are based on information compression [38] and other disturbance mechanisms to achieve local differential privacy. RR is initially designed for sensitive questions with binary answers "yes" or "no". Users can choose whether to upload the original answer or the reversed response depending on coin-flipping. RR is then easily extended to make statistics on categorical data for frequency estimation. RAPPOR [23] developed by Google. inc encodes data as a Bloom filter and makes the randomized response on each Bloom filter bit. RAPPOP realized privacy-preserving data aggregation and analysis for data categories, frequencies, and other set statistics in crowdsourcing. Wang et al. [24] proposed a protocol for finding frequent items in the set-valued LDP setting. Ren et al. [39] focused on the high-dimensional crowdsourced data publication with a guarantee of LDP. In Ref. [22], discrete distribution estimation based on the k-subset mechanism satisfying LDP has been proposed. In Ref. [40], the author focused on the frequency and mean estimation for key-value type data with LDP. Most of the works on LDP focus on frequency estimation and distribution estimation for discrete categorical data. In this paper, we propose a demand estimation scheme for the LDP-perturbed numerical data that is aggregated in a smart grid.

3. Problem model

3.1. Network model

We simplify the network model of data acquisition in the smart grid. The network model consists of three kinds of entities in our system, Smart Meter (SM), Gateway (AG), and Control Center (CC). The whole organization of the model is shown in Fig. 1. What follows are their functions and roles.

- **Smart Meter (SM):** SM is an intelligent device that is installed in each user's house. It has limited computing power, thus, the computation burden on the meter side should be as small as possible. The smart meter collects and submits the power consumption data of a single user. For clarity, we treat the "smart meter" and "user" the same and may use them interchangeably.
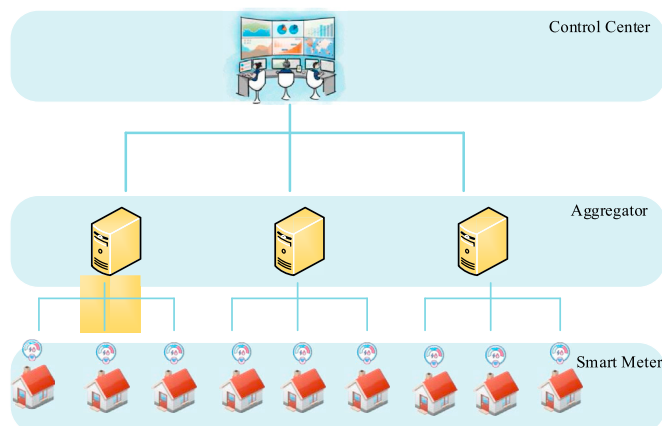

Fig. 1. System model.

- **Gateway (AG):** The gateway acts as an aggregator in the smart grid system. It collects and aggregates data submitted from smart meters. After aggregation, it sends the aggregated result to the control center for further analysis.
- **Control Center (CC):** CC connects with all AGs and collects the data of total power consumption from AGs. Then, CC formulates a power dispatching strategy and adjusts electricity prices.

The smart meter installed in each house reports the electricity consumption data to the control center periodically in the smart grid. The aggregator collects and aggregates data submitted from smart meters. After aggregation, it sends the aggregated result to the control center. The control center gathers all the data submitted from smart meters, performs statistical analysis, and then manages the smart grid's electricity generation, transmission, and distribution.

3.2. Security assumption

Electric supply companies manage the CC and AG in the smart grid. Therefore, the CC and AG can be called honest but curious. They will process the data in accordance with the data aggregation methodology, but they are also concerned about user data privacy. When the control center and the gateway perform data statistics and analysis duties on the smart grid, there should be a plan in place to ensure that the original data provided by users is not obtained.

Users are treated as honest participants who process and submit generated power consumption data to the gateways in accordance with the protocol; nonetheless, they are concerned about data privacy.

In addition, we consider attackers who acquire and observe data uploaded by users in our system. The attackers who intercept and tamper with the user's data are not taken into account. We believe that secure communication channels exist between users and aggregators.

3.3. Security and design goals

Based on the above assumption of each entity in the smart grid system, our scheme is proposed to achieve the following goals:

- *Data privacy.* The data collected by aggregator may disclose users' privacy. Therefore, during the data aggregation, the privacy of users' data should be preserved. The aggregator should know nothing about any particular user's data but the final aggregation result. It is worth noting that data aggregation tasks in smart grids are periodic and frequent, so it is also necessary to ensure that the privacy of user data will not be disclosed in the long-term collection process.
- *Practicability.* Data is submitted on a regular and frequent basis in the smart grid. As a result, data processing and submission efficiency become critical. The calculation overhead of each smart meter should be manageable because the smart meter does not have a lot of computing power. Furthermore, the communication overhead between users should be kept to a minimum.
- *No need for a trusted third party.* Users tend to be skeptical that the entities who have access to their data will threaten their data privacy, and in the real world, it is unrealistic to assume a trusted third party. The assumption of a trusted third party is not realistic in practice. Therefore, our proposed scheme should not rely on a trusted third party.
- *Support dynamic changes of users.* In a smart grid, users may join or quit the system, so the aggregation scheme should accommodate the dynamic change of users. One data aggregation task in a smart grid often requires thousands of users to participate. Therefore, when the users change dynamically, the system's extra computation and communication cost should be as small as possible. More specifically, when some users join or leave the system, the other users in the system do not need to renegotiate new parameters.

4. Preliminaries

4.1. Local Differential Privacy (LDP)

The formal definition of local differential privacy is as follows:

Definition 1. For any user i, an algorithm $\mathcal{M}$ satisfies ε-local differential privacy (ε-LDP), where $\varepsilon \geq 0$, if and only if for any two data records X^i, X^j, and for any possible outputs $\tilde{X} \in Range(\mathcal{M})$,

$$Pr[\mathcal{M}(X^i) = \tilde{X}] \leqslant e^\varepsilon \times Pr[\mathcal{M}(X^j) = \tilde{X}]$$

where value ε is called the privacy budget. It can be seen that LDP ensures that algorithm $\mathcal{M}$ satisfies ε-LDP by controlling the similarity of the output results of any two records. In a nutshell, the adversary seeing $\tilde{X}$ cannot determine whether the input is X^i or X^j. For more details about local differential privacy, refer to the introduction and summary of LDP in Ref. [35].

4.2. k-Randomized Response (k-RR)

The k-Randomized Response (k-RR) is a typical randomized response scheme for aggregating and analyzing discrete categorical data. The perturbation function is defined as: For any input $R \in \mathcal{X}$ and its corresponding output $R' \in \mathcal{X}$, there exists

$$P(R'|R) = \begin{cases} p = \dfrac{e^\varepsilon}{k - 1 + e^\varepsilon}, & R' = R \\[2mm] q = \dfrac{1}{k - 1 + e^\varepsilon}, & R' \neq R \end{cases}$$

where ε is the privacy budget and $k = |\mathcal{X}|$.

The k-RR satisfies ε-LDP since for any inputs R_1, R_2 and output R', there exists

$$\frac{P(R'|R_1)}{P(R'|R_2)} \leq \frac{p}{q} = e^\varepsilon$$

To estimate the frequency of $R \in \mathcal{X}$, the aggregator counts how many times R is submitted as $C(R)$, and then computes

$$\Phi(R) = \frac{C(R) - nq}{p - q}$$

where n is the total number of the users, $\Phi(R)$ is the estimation of the number of users whose input value is R.

As analyzed in Ref. [37], the estimation variance is

$$Var(\Phi(R)) = n \cdot \frac{k - 2 + e^\varepsilon}{(e^\varepsilon - 1)^2}$$

the variance grows while the number of categories becomes large. When the number of the category is not large, the k-RR has good performance in frequency estimation.

References [36,41] provide more details about k-Randomized Response (k-RR).

5. Scheme description

Before we describe our scheme in detail, we first present the overview of our scheme, which shows the core techniques and functional features.

5.1. Overview

The CC must obtain statistical data on power use over time in order to estimate the smart grid's power consumption. Our goal is to design a smart grid data aggregation scheme that is efficient and practical. The main idea is to discretize data and estimate the total or average power consumption by analyzing data frequency through RR. However, a straightforward combination of data discretization and RR will make the scheme lose great data precision. To increase aggregation accuracy, we propose a special data discretization scheme to reduce the accuracy loss. We develop a data aggregation approach that meets LDP for numerical data by combining it with a randomized answer.

In our scheme, the smart meter first transforms the generated data according to a specific probability, which is dependent on the generated data, before the operation of a randomized response. Specifically, the actual generated data are first converted into discrete values. Then RR is performed on the transformed discrete data. Then, the aggregator collects and analyzes the data submitted from users in the system, and estimates the frequency of each discrete value. Finally, the aggregator gets the statistical results and completes the demand estimation of the smart grid. Fig. 2 shows the process of the data aggregation scheme.

Note that the discrete interval division in our scheme needs not to be the same, and the aggregator can decide it according to the data analysis demand.

5.2. System initialization phase

In the system initialization phase, the control center first determines the parameters such as the legal range of data uploaded by users, the privacy budget ε, and the time interval of data collection tasks.

Since power consumption data that are generated by smart meters are always in a certain range, it is reasonable to assume that the raw data submitted by honest users are within the interval $[0, m]$. Then the CC divides the interval into $[0, s)$, $[s, 2s)$, ..., $[(d-1)s, ds]$, assuming that $d = \left\lceil \frac{m}{s} \right\rceil$. For the sake of presentation clarity, the interval is split evenly into several subintervals. In practice, the CC can divide the intervals into arbitrary lengths according to the demand. Note that ds will be larger than m when m cannot be divisible by s, but it will not influence the correctness of our scheme. We record the set of boundary values of all subintervals as X and the number of natural number elements in a subinterval as $|X|$. In this case, $X = \{0, s, 2s, ..., ds\}$ and $|X| = d + 1$.

Then, the CC broadcasts the interval $[0, m]$, the subintervals $[0, s)$, $[s, 2s)$, ..., $[(d-1)s, ds]$ and the privacy budget ε to gateways and all of the smart meters in the system.

Algorithm 1. Data perturbation algorithm

Algorithm 1 Data perturbation algorithm

Input: The division $[0, s)$, $[s, 2s)$, ..., $[(d-1)s, ds]$ of the data range $[0, m]$. Raw data $x_i \in [0, m]$ generated by smart meter u_i; $x_i \in [u, v], u, v \in X$. Privacy parameter ϵ.

Output: Data $y_i \in X$ submitted to the aggregator.

1: Discretization:

$$p(x_i'|x_i) = \begin{cases} \dfrac{v - x_i}{v - u}, & x_i' = u \\[2mm] \dfrac{x_i - u}{v - u}, & x_i' = v \end{cases}$$

2: Perturbation:

$$p(y_i|x_i') = \begin{cases} \dfrac{e^\epsilon}{|X| - 1 + e^\epsilon}, & y_i = x_i' \\[3mm] \dfrac{1}{|X| - 1 + e^\epsilon}, & y_i \neq x_i' \end{cases}$$
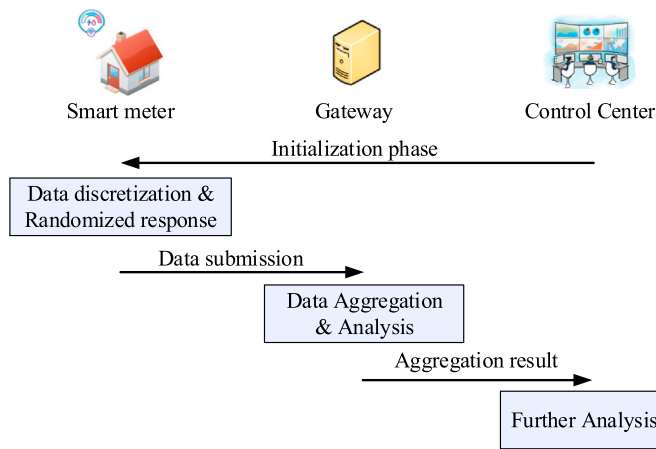
return y_i;

Fig. 2. Process of the data aggregation scheme.

5.3. Data submission phase

For a user u_i with power consumption data x_i that belongs to the subinterval $[\lfloor \frac{x_i}{s} \rfloor \cdot s, (\lfloor \frac{x_i}{s} \rfloor + 1) \cdot s)$, user u_i generates the data to be submitted to the gateway according to the following steps. For the convenience of scheme description, we express this subinterval as $[u, v]$, where $u = \lfloor \frac{x_i}{s} \rfloor \cdot s$ and $v = (\lfloor \frac{x_i}{s} \rfloor + 1) \cdot s$.

1. First, user u_i discretizes the actual numerical data x_i that smart meter generated to a natural number x_i' with the conditional probability $p(x_i'|x_i)$, which is computed according to the value x_i as follows:

$$p(x_i'|x_i) = \begin{cases} \dfrac{v - x_i}{v - u}, & x_i' = u \\ \dfrac{x_i - u}{v - u}, & x_i' = v \end{cases}$$

We can see that x_i is discretized to the boundary value of the interval to which x_i belongs.

2. Then, user u_i uses the discretized value x_i' to generate the submitted data by k-RR with a certain probability. We consider the final result calculated by u_i as $y_i \in X$, of which the corresponding probability is as follows:

$$p(y_i|x_i') = \begin{cases} p = \dfrac{e^\varepsilon}{|X| - 1 + e^\varepsilon}, & y_i = x_i' \\ q = \dfrac{1}{|X| - 1 + e^\varepsilon}, & y_i \neq x_i', y_i \in X \end{cases}$$

3. Finally, user u_i (smart meter) submits the perturbed result y_i to the gateway for data aggregation.

5.4. Data aggregation and analysis

After receiving y_i from all users in its administrative region, the gateway aggregates and analyzes these data. Since each $y_i \in X$, the gateway can get the total power consumption by counting the frequency of each element in X. We denote the frequency of each element $X_j \in X$, $0 \leq j \leq |X|$ as $C(X_j)$. The gateway computes

$$\Phi(X_j) = \frac{C(X_j)(|X| - 1 + e^\varepsilon) - n}{e^\varepsilon - 1}$$

$\Phi(X_j)$ is the estimated value for the frequency of each element X_j.

Then, the gateway can estimate the total power consumption as

$$R = \sum_{j=1}^{j=|X|} y_j \cdot \Phi(y_j)$$

Then, the gateways in the system send the total power consumption data R and the estimated value $\Phi(X_j)$ for the frequency of each element X_j to the CC, which can then get statistical information by analyzing, such as the average and peak of the power consumption.

In addition to aggregating power consumption data, CC can require gateways to perform other statistical analyses of the power consumption in the smart grid such as mode analyses. CC can get a wealth of power consumption information for improving its services.

6. Further discussion of the aggregation scheme

In Section 5, we introduce our data aggregation scheme in detail. The data aggregation scheme we proposed can fit the smart grid's general supply and demand estimation scenario. When the data range becomes large and the number of subintervals increases greatly, as analyzed in Ref. [37], the accuracy of the k-RR deteriorates when the number of subintervals increases. For example, when aggregating the power consumption data of industrial power consumption in smart grid, the value range of power consumption data is much larger than that of ordinary residents. In this case, the consideration for more subintervals is necessary.

To solve the problem that the accuracy of the scheme decreases when the number of subintervals increases, we propose an extended solution. As analyzed in Ref. [37], when the number of subintervals, specifically, when $|x| < 3e^\varepsilon + 2$, the result of k-RR is acceptable. We consider using the grouping method to divide the whole data value range.

In the system initialization phase, the control center first determines the privacy budget ε, the CC determines the appropriate number of subintervals d according to the selection of privacy budget where $d < 3e^\varepsilon + 2$ and the subinterval length s. Then, using the number of subintervals d and the subinterval length s to divide the whole data range $[0, m]$ in to several groups, the group quantity is $t = \lceil \frac{m}{d \cdot s} \rceil$. Mark each group as g_k, $1 \leq k \leq t$. Then, the CC broadcasts the interval $[0, m]$, the privacy budget ε, and the division of groups and subintervals to gateways and all of the smart meters in the system.

During the data submission phase, the user u_i first determines the group to which the generated data belongs $g_i = \lfloor \frac{x_i}{d \cdot s} \rfloor$. The sub interval within the group is $[\lfloor \frac{x_i - g_i \cdot ds}{s} \rfloor \cdot s, (\lfloor \frac{x_i - g_i \cdot ds}{s} \rfloor + 1) \cdot s)$. In this case, $u = \lfloor \frac{x_i - g_i \cdot ds}{s} \rfloor \cdot s$ and $v = (\lfloor \frac{x_i - g_i \cdot ds}{s} \rfloor + 1) \cdot s$. Then, the user perturbs the data x_i according to the data perturbation algorithm proposed according to Algorithm 1. After perturbing the data, user u_i submits the perturbed result y_i and the group number g_i to which the data belongs to the gateway for data aggregation.

The gateway collects all data submitted by smart meters in the system and does the aggregation. For each group g_k, $1 \leq k \leq t$, the aggregator estimates the frequency of data in each group separately. We mark the j-th element of group k as X_{kj}, $1 \leq k \leq t$, $0 \leq j \leq d$. The frequency of each element is $C(X_{kj})$. Then, the gateway computes

$$\Phi(X_{kj}) = \frac{C(X_{kj})(d - 1 + e^\varepsilon) - n_{g_k}}{e^\varepsilon - 1}$$

The total power consumption is

$$R = \sum_{k=1}^{k=t} \sum_{j=1}^{j=d} y_{kj} \cdot \Phi(y_{kj})$$

In this case, for the convenience of description, we still divide the whole data interval equally. In practical application, the control center can select the interval that is not evenly divided according to the situation.

7. Privacy and utility analysis

In this section, we analyze and evaluate the privacy and utility of our scheme, and prove that our scheme meets the proposed design goals. We first consider the privacy protection and accuracy of our scheme. Then, we analyze how our scheme supports the dynamic changing of users. Finally, we give the discussion for a situation of unevenly distributed intervals.

7.1. Privacy analysis

Theorem 1. *The proposed smart meter data processing scheme satisfies ε-local differential privacy.*

Proof. In our scheme, the process of generating y_j from $x_i^{'}$ satisfies the k-Randomized Response given in Section 4. *Assuming that any two elements x_i, x_j in X, we can have*

$$\frac{Pr(x_i|x)}{Pr(x_j|x)} \leq \frac{\frac{e^\varepsilon}{|X|-1+e^\varepsilon}}{\frac{1}{|X|-1+e^\varepsilon}} = e^\varepsilon$$

In our scheme, the final submitted data y_i satisfies

$$P(y_i|x_i) = \begin{cases} \frac{v-x_i}{v-u} \cdot p + \frac{x_i-u}{v-u} \cdot q, & y_i = u \\ \frac{x_i-u}{v-u} \cdot p + \frac{v-x_i}{v-u} \cdot q, & y_i = v \\ q, & y_i \neq u, v \end{cases}$$

We assume that $x_i - u \leq v - x_i$, and there is

$$\frac{P(y_i|x_i)}{P(y_i^{'}|x_i)} \leq \frac{\frac{v-x_i}{v-u} \cdot \frac{e^\varepsilon}{|X|-1+e^\varepsilon} + \frac{x_i-u}{v-u} \cdot \frac{1}{|X|-1+e^\varepsilon}}{\frac{1}{|X|-1+e^\varepsilon}}$$

$$= \frac{v-x_i}{v-u} \cdot e^\varepsilon + \frac{x_i-u}{v-u} \leq e^\varepsilon$$

It is easy to prove that when $v - x_i \leq x_i - u$, $\frac{P(y_i|x_i)}{P(y_i^{'}|x_i)} \leq e^\varepsilon$ also establishes. Therefore, the proposed scheme satisfies ε-local differential privacy.

The data collected by the control center are locally perturbed and satisfy LDP. Therefore, CC can only aggregate and analyze the data submitted by all smart meters to obtain the statistical results of the data, but cannot know the original data content of a single user. Overall, data privacy can be guaranteed in our scheme. Moreover, due to the user's special discretization before a randomized response to the data, the same data has the probability of being converted to different outputs. During the long-term collection of user data, the user's data privacy will not be destroyed.

7.2. Accuracy analysis

Theorem 2. *The data discretization in our scheme does not reduce the statistical accuracy of data.*

Proof. When user u_i discretizes the raw data $x_i \in [u, v)$, $u, v \in X$, he gets u with probability $\frac{v-x_i}{v-u}$ and gets v with probability $\frac{x_i-u}{v-u}$, therefore, the expectation of x_i is

$$E(x_i^{'}) = u \cdot \frac{v-x_i}{v-u} + v \cdot \frac{x_i-u}{v-u} = x_i$$

The expectation of $x_i^{'}$ is equal to x_i, thus, there is no accuracy loss in the process of discretization.

When the users' data are not uniformly distributed, the expectation of $x_i^{'}$ is not equal to x_i if the data is simply discretized instead of adopting our proposed algorithm, which increases deviation between the data aggregation result and actual aggregation result.

7.3. Support for user dynamics

The join and exit of smart meters in the smart grid may participate in data aggregation. Each smart meter in the system can conduct discretization and perturbation locally with the input: data range, the division of subintervals, and the privacy budget, making it very convenient for new users to join the system.

When there exist users exiting the system or failing to submit data, as long as CC collects enough amount of data from other smart meters in the smart grid, it can still aggregate and analyze the data normally to obtain the estimation of power supply and demand of the smart grid. Therefore, our scheme has good support for users' joining and exiting in the smart grid.

7.4. Situation of unevenly distributed interval

In Section 5, we divide the interval of submitted data equally to explain the scheme's content more clearly. While in practice, the interval can be distributed into uneven subintervals. The AG can reduce the interval near the average power consumption of users according to the experience. The data range [0, m] can be divided into $[0, m_1)[m_1, m_2)…[m_j, m_j + 1), …, [md_1), m]$, where the size of each subinterval can be different. This will not threaten users' data privacy or reduce the utility of aggregation results, and at the same time, CC can analyze the general power consumption habits of users in the region to a certain extent.

8. Performance evaluation

This section will analyze our scheme's accuracy and efficiency through comprehensive measurements. We first analyze the accuracy performance with different privacy budgets ε and the number of subintervals s. Then, we analyze the utility of the extended scheme in the case of the large data range. Then, we compare the scheme proposed in Section 5 with two typical data aggregation schemes [13,18] in the smart grid in terms of computation overhead and communication overhead. In Ref. [13], smart meters' data privacy protection is realized by generating and distributing random numbers for each smart meter in advance. In Ref. [18], the smart meters add noise to raw data and upload it to the gateway through homomorphic encryption, and then the gateway obtains the final result through calculation on the ciphertext.

The experiments below are implemented on a standard 64-bit Windows 10 system with a 3.00 GHz Intel Core i5 processor. Our scheme is implemented by Python (with version 3.7). The homomorphic encryption we use is Paillier encryption from the phe library[1] of Python. If there are no additional statements in the experiments, the number of users we set is 1,000 in an aggregation task, and the submitted data range from 0 to 100 divided into 10 subintervals.

8.1. Utility analysis

To assess the value of our scheme's statistical outcomes, we consider different privacy budgets ε and numbers of subintervals s. We produce 1,000 numbers at random in the range [0,100] to simulate 1,000 user data in data aggregation jobs with an accurate aggregation result of 48,155. As shown in Fig. 3 and Fig. 4, the circle dotted line represents the actual value of the data aggregation task.

We first evaluate the impact of privacy budgets on statistical analysis, and Fig. 3 shows the results under different privacy budgets. With the same number of subintervals s, the larger the privacy budget ε is, the closer the statistical results are to the real value. This result is consistent with the expectation of theoretical analysis. When the privacy budget ε is small, the probability of the submitted data falling into other intervals is much greater than that of the original interval, which makes the

---

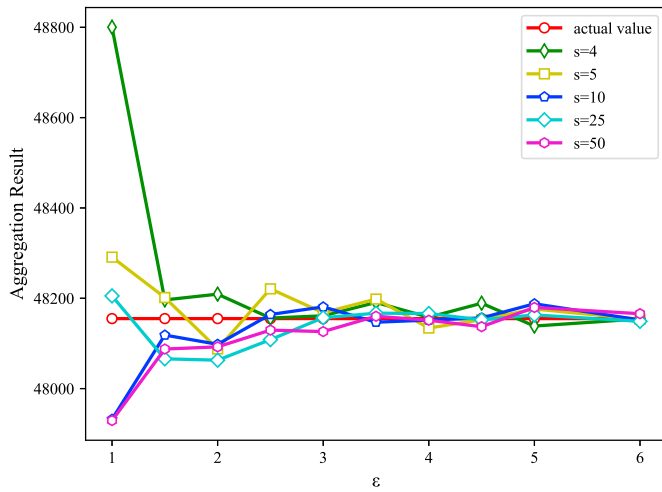[1] https://pypi.org/project/phe/1.0/.

Fig. 3. Statistical result with the relationship of privacy budget and data accuracy.

estimated value largely deviate from the real value. When the privacy budget is larger than 1.5, the error of the statistical results obtained by our scheme is relatively small, and the data accuracy is guaranteed. In the real world, the gateway can set different privacy budgets for different situations, so as to ensure the privacy of users' data in different degrees under the condition of data utility.

Then, we test the relationship between subinterval number and data utility, and Fig. 4 shows the results with different subinterval numbers *s*. When $\varepsilon = 1$, the results fluctuate as the subinterval number grows. When the value of the privacy budget is reasonable (larger than 1.5 according to the aforementioned experiment), the results keep steady with acceptable errors. Thus, the subinterval number is not the main factor that influences the utility performance.

8.2. Utility analysis for the situation of large range of values

In this section, we analyze the scheme's utility in the case of a large data range. We randomly generate 10,000 numbers in [0,1000] to simulate ten thousand user data in data aggregation tasks, the privacy budget ε is set as 2, and the correct mean value is 497.555. In this case, the number of subintervals becomes large without the proposed grouping method. As shown in Fig. 5 and Fig. 6, the circle dotted line represents the actual value of the data aggregation task.
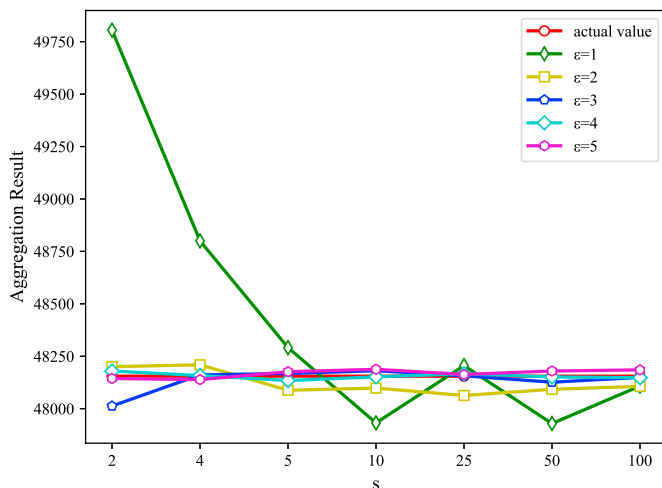
Fig. 5 shows the actual mean value and the estimated mean value without the grouping method. The line with a diamond-shaped dot shows the mean estimation under different subinterval numbers. As the number of subintervals increases, the utility of the aggregation result reduces. It can be seen in the figure when the number of subintervals is more than 50. The estimation error becomes larger and it is consistent with the theoretical analysis, as in our simulation $3e^{\varepsilon} + 2 = 24.17$. Under normal circumstances, in smart grid data acquisition scenarios, the data range will not be too large, the number of subintervals can be kept in the appropriate range, and the data utility is guaranteed. As the number of subintervals increases, the user's data is disturbed into other intervals with a greater probability. When there are multiple subintervals, the error between the statistical results and the real value will increase. As shown in Fig. 5, when the number of intervals is larger than 40, the error of statistical results increases, and the utility of statistical results decreases. Consequently, if the data range is vast, we need to consider the problem that the error increases when the number of subintervals increases.

Fig. 6 shows the comparison between estimation with the grouping method and without the grouping method. We make experiments and analyses on the cases of 5, 10, and 20 groups. With the reduction in the length of subintervals, the number of subintervals in each group increases. Experimental results show that when the data range is large, the grouping method can greatly improve the utility of data aggregation results. Because of the adoption of grouping, the number of subintervals in each group can remain in an appropriate range, which reduces the error of the estimated value. After grouping, the estimation error of the mean value is greatly reduced, and the data utility is guaranteed when the length of subintervals decreases.

8.3. Computation overhead

In this section, we analyze and compare our scheme with Gope's scheme [13] and Bao's scheme [18].

Table 1 shows the comparison result of the computation overhead of smart meters and aggregators (gateways in our scheme). Smart meters and aggregators only need to conduct real number addition and multiplication several times, which leads to little computation overhead. In Gope's Scheme, smart meters will conduct one hash operation to preserve privacy, and aggregators are required to conduct *n* hash operations and symmetric encryptions. Bao's scheme includes some exponentiation and multiplication over a cyclic group *G*. Fig. 7 shows experimental results. The overhead of exponentiation and multiplication over *G* are much larger than other operations, Bao's scheme consumes the longest time
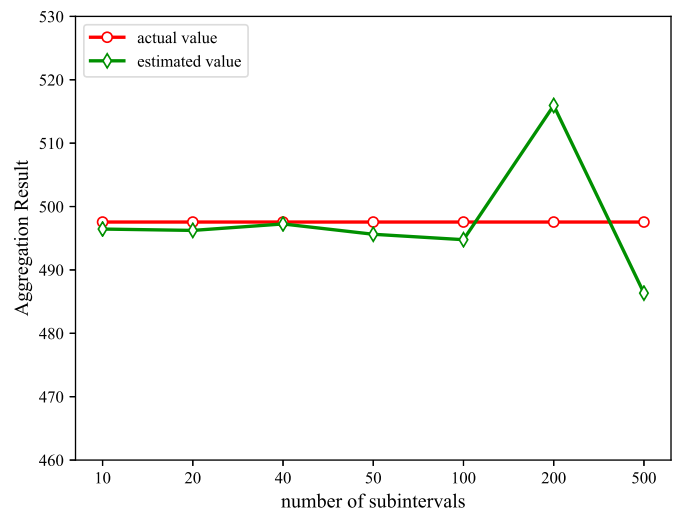

Fig. 4. Statistical result with the relationship of subinterval number and data accuracy.

Fig. 5. Statistical result with the relationship of subinterval quantity and data utility without grouping.
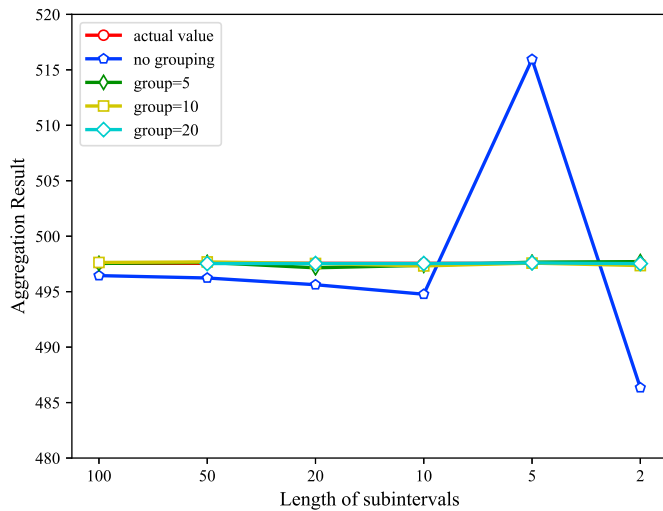
Fig. 6. Statistical result with the relationship of subinterval quantity and data utility comparison.

Table 1
Computation overhead comparison.

Scheme	Operations	
	Smart Meter	Aggregator
Gope's Scheme [13]	H	$n(SE + H)$
Bao's Scheme [18]	$2^a C_e + C_m$	$(n-1)C_m$
Our Scheme	$3ADD + 1MUL$	$|X|(4ADD + 3MUL)$

[a] We denote hash, symmetric encryption, exponentiation over a cyclic group G, multiplication over G, number of user, real number addition and real number multiplication as H, SE, C_e, C_m, n, ADD, MUL, respectively.

during the aggregation process. However, benefitting from the light-weight real number addition and multiplication, our scheme has the best performance, with 0.022 5 ms time cost at smart meters and 0.045 4 ms time cost at aggregators.

Fig. 8 shows the overall computation overhead of the aggregators with varying numbers of smart meters participating in the data aggregation process. When the number of smart meters participating in the task is not large, the computation costs of the three schemes are all acceptable. However, the computation overhead of Gope's scheme and Bao's scheme will grow approximately linearly as the number of participating meters grows. While the computation overhead of the
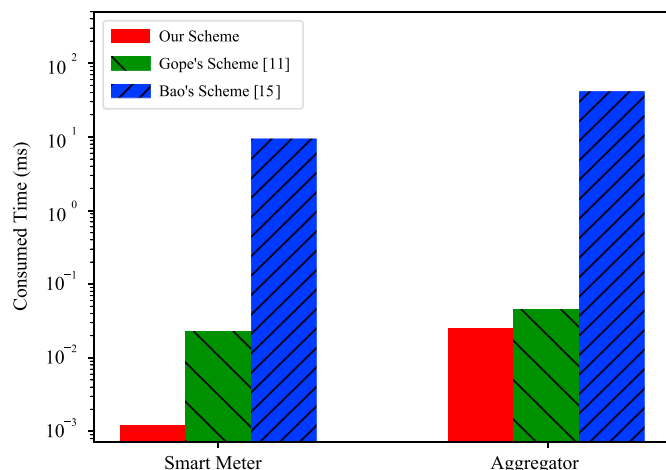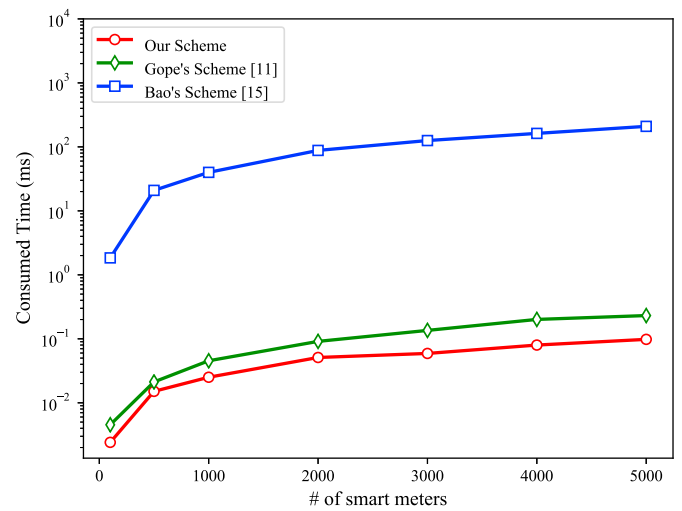

Fig. 8. Computation overhead of the aggregators with varying numbers of smart meters.

aggregators remains stable, the reason is that the aggregators only have to count the number of each discrete value and calculate the final result. Consequently, the computation overhead is much smaller than schemes based on cryptography algorithm.

In the real-world smart grid system, the number of smart meters participating in data aggregation tasks is often very large. Moreover, the tasks of data aggregation are often very frequent in the smart grid, which requires the computation overhead of every single data aggregation task to be as small as possible. By analyzing computing overhead, we can conclude that our scheme is more efficient and practical than other schemes in the real-world smart grid system.

8.4. Communication overhead

In this section, we look at our scheme's communication overhead and compare it to two others. Fig. 9 shows the entire communication overhead of a single aggregating process with a various numbers of participating smart meters. Because of the vast range of communication overhead, we use logarithmic coordinates to show experimental results.

For fairness consideration, in Gope's scheme, we only consider the communication cost in the data aggregation part of the scheme. As shown in Fig. 9, the communication overhead of these three schemes increases with the increasing number of smart meters. Because there is no
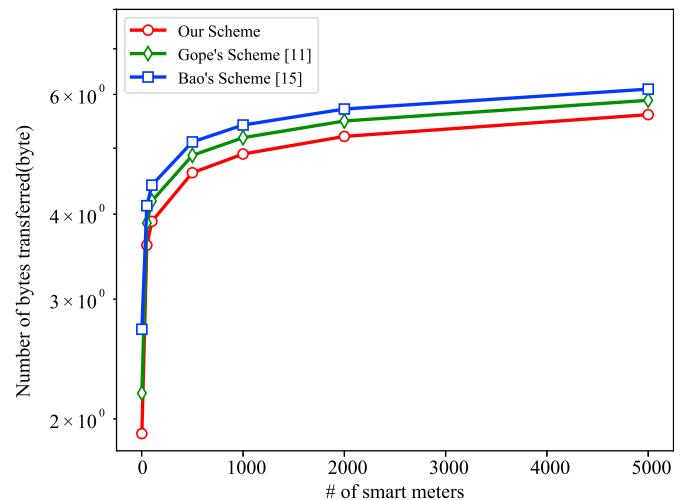

Fig. 7. Computation overhead of different entities.

Fig. 9. Communication overhead with varying numbers of smart meters.

encryption in our scheme, only the division of the interval, the privacy budget, perturbed data, and aggregation results must be transmitted during the aggregation process. Considering the large volume of ciphertext, our scheme has the lowest communication overhead. While in Gope's scheme, besides necessary perturbed data and aggregation results, it needs to transmit two extra hash values and one ciphertext. Thus, the communication overhead of Gope's scheme is larger than ours. Bao's scheme introduces encryption based on a cyclic group G, of which the overall communication overhead is $(2n + 2) \cdot L_G$, leading to the largest communication overhead. Here L_G is the output of the modular operation in G assumed to be 1024 bit.

9. Conclusions

In this study, we proposed a privacy-preserving data aggregation scheme for the smart grid. Considering the limited computation ability of smart meters, we reduced the computational burden of smart meters that participate in data aggregation tasks. By designing a special data discretization algorithm and random response mechanism, the scheme achieves the privacy-preserving smart grid data aggregation satisfying the LDP. Furthermore, we consider more special scenarios, which enable our scheme to cope with the special situation of large data range in the smart grid data aggregation scenario. Unlike existing schemes based on masking values, our scheme can run normally without a trusted third party. Users need not negotiate for the masking values, and our scheme can also deal with users' joining and exiting in the smart grid. Through the comprehensive analysis, our scheme is shown to be privacy-preserving with less computation and communication overhead compared with other available literature.

Declaration of competing interest

The authors declared that they have no conflicts of interest in this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] V.C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, G.P. Hancke, Smart grid technologies: communication technologies and standards, IEEE Trans. Ind. Inf. 7 (4) (2011) 529–539.

[2] A. Abdallah, X.S. Shen, A lightweight lattice-based homomorphic privacy-preserving data aggregation scheme for smart grid, IEEE Trans. Smart Grid 9 (1) (2016) 396–405.

[3] Y. Liu, W. Guo, C.-I. Fan, L. Chang, C. Cheng, A practical privacy-preserving data aggregation (3PDA) scheme for smart grid, IEEE Trans. Ind. Inf. 15 (3) (2018) 1767–1774.

[4] S. Li, X. Zhang, K. Xue, L. Zhou, H. Yue, Privacy-preserving prepayment based power request and trading in smart grid, China Communications 15 (4) (2018) 14–27.

[5] H. Khurana, M. Hadley, N. Lu, D.A. Frincke, Smart-grid security issues, IEEE Security & Privacy 8 (1) (2010) 81–85.

[6] P. McDaniel, S. McLaughlin, Security and privacy challenges in the smart grid, IEEE Security & Privacy 7 (3) (2009) 75–77.

[7] S. Li, K. Xue, D.S. Wei, H. Yue, N. Yu, P. Hong, SecGrid: a secure and efficient SGX-enabled smart grid system with rich functionalities, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1318–1330.

[8] L. Chen, R. Lu, Z. Cao, PDAFT: a privacy-preserving data aggregation scheme with fault tolerance for smart grid communications, Peer-to-Peer.Netw. Appl. 8 (6) (2015) 1122–1132.

[9] T.W. Chim, S.-M. Yiu, V.O. Li, L.C. Hui, J. Zhong, PRGA: privacy-preserving recording & gateway-assisted authentication of power usage information for smart grid, IEEE Trans. Dependable Secure Comput. 12 (1) (2014) 85–97.

[10] F. Li, B. Luo, P. Liu, Secure information aggregation for smart grids using homomorphic encryption, in: Proceedings of the First IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2010, pp. 327–332.

[11] R. Lu, X. Liang, X. Li, X. Lin, X. Shen, EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications, IEEE Trans. Parallel Distr. Syst. 23 (9) (2012) 1621–1631.

[12] K. Xue, Q. Yang, S. Li, D.S. Wei, M. Peng, I. Memon, P. Hong, PPSO: a privacy-preserving service outsourcing scheme for real-time pricing demand response in smart grid, IEEE Internet of Things J. 6 (2) (2018) 2486–2496.

[13] P. Gope, B. Sikdar, An efficient data aggregation scheme for privacy-friendly dynamic pricing-based billing and demand-response management in smart grids, IEEE Internet Things J. 5 (4) (2018) 3126–3135.

[14] X. Gong, Q.-S. Hua, L. Qian, D. Yu, H. Jin, Communication-efficient and privacy-preserving data aggregation without trusted authority, in: Proceedings of the 2018 IEEE Conference on Computer Communications (INFOCOM), IEEE, 2018, pp. 1250–1258.

[15] W. Jia, H. Zhu, Z. Cao, X. Dong, C. Xiao, Human-factor-aware privacy-preserving aggregation in smart grid, IEEE. Syst.J 8 (2) (2013) 598–607.

[16] K. Xue, B. Zhu, Q. Yang, D.S. Wei, M. Guizani, An efficient and robust data aggregation scheme without a trusted authority for smart grid, IEEE Internet Things J. 7 (3) (2019) 1949–1959.

[17] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, Our data, ourselves: privacy via distributed noise generation, in: Proceedings of the 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt), Springer, 2006, pp. 486–503.

[18] H. Bao, R. Lu, DDPFT: secure data aggregation scheme with differential privacy and fault tolerance, in: Proceedings of the 2015 IEEE International Conference on Communications (ICC), IEEE, 2015, pp. 7240–7245.

[19] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Proceedings of the 2006 Theory of Cryptography Conference (TCC), Springer, 2006, pp. 265–284.

[20] J.C. Duchi, M.I. Jordan, M.J. Wainwright, Local privacy and statistical minimax rates, in: Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2013, pp. 1592–1597.

[21] C. Xu, J. Ren, D. Zhang, Y. Zhang, Distilling at the edge: a local differential privacy obfuscation framework for iot data analytics, IEEE Commun. Mag. 56 (8) (2018) 20–25.

[22] S. Wang, L. Huang, Y. Nie, X. Zhang, P. Wang, H. Xu, W. Yang, Local differential private data aggregation for discrete distribution estimation, IEEE Trans. Parallel Distr. Syst. 30 (9) (2019) 2046–2059.

[23] Ú. Erlingsson, V. Pihur, A. Korolova, RAPPOR: randomized aggregatable privacy-preserving ordinal response, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2014, pp. 1054–1067.

[24] T. Wang, N. Li, S. Jha, Locally differentially private frequent itemset mining, in: Proceedings of the 2018 IEEE Symposium on Security and Privacy (S&P), IEEE, 2018, pp. 127–143.

[25] J. Liu, C. Zhang, Y. Fang, EPIC: a differential privacy framework to defend smart homes against internet traffic analysis, IEEE Internet Things J. 5 (2) (2018) 1206–1217.

[26] N. Gai, K. Xue, P. He, B. Zhu, J. Liu, D. He, An efficient data aggregation scheme with local differential privacy in smart grid, in: Proceedings of the 16th International Conference on Mobility, Sensing and Networking (MSN), IEEE, 2020, pp. 73–80.

[27] Y. Ding, B. Wang, Y. Zhang, K. Zhang, H. Wang, Secure metering data aggregation with batch verification in industrial smart grid, IEEE Trans. Ind. Inf. 16 (10) (2020) 6607–6616.

[28] S. Zhao, F. Li, H. Li, R. Lu, S. Ren, H. Bao, J.-H. Lin, S. Han, Smart and practical privacy-preserving data aggregation for fog-based smart grids, IEEE Trans. Inf. Forensics Secur. 16 (2020) 521–536.

[29] P. Paillier, Public-key cryptosystems based on composite degree residuosity classes, in: Proceedings of the 17th International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), Springer, 1999, pp. 223–238.

[30] S. Li, K. Xue, Q. Yang, P. Hong, PPMA: privacy-preserving multisubset data aggregation in smart grid, IEEE Trans. Ind. Inf. 14 (2) (2017) 462–471.

[31] D. Boneh, E.-J. Goh, K. Nissim, Evaluating 2-DNF formulas on ciphertexts, in: Proceedings of 2005 Theory of Cryptography Conference (TCC), Springer, 2005, pp. 325–341.

[32] H. Bao, R. Lu, A new differentially private data aggregation with fault tolerance for smart grid communications, IEEE Internet Things J. 2 (3) (2015) 248–258.

[33] C. Castelluccia, A.C. Chan, E. Mykletun, G. Tsudik, Efficient and provably secure aggregation of encrypted data in wireless sensor networks, ACM Trans. Sens. Netw. 5 (3) (2009) 1–36.

[34] L. Lyu, K. Nandakumar, B. Rubinstein, J. Jin, J. Bedo, M. Palaniswami, PPFA: privacy preserving fog-enabled aggregation in smart grid, IEEE Trans. Ind. Inf. 14 (8) (2018) 3733–3744.

[35] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, T. Wang, Privacy at scale: local differential privacy in practice, in: Proceedings of 2018 International Conference on Management of Data (SIGMOD), ACM, 2018, pp. 1655–1658.

[36] S.L. Warner, Randomized response: a survey technique for eliminating evasive answer bias, J. Am. Stat. Assoc. 60 (309) (1965) 63–69.

[37] T. Wang, J. Blocki, N. Li, S. Jha, Locally differentially private protocols for frequency estimation, in: Proceedings of the 26th USENIX Security Symposium (USENIX Security), USENIX Association, 2017, pp. 729–745.

[38] S. Xiong, A.D. Sarwate, N.B. Mandayam, Randomized requantization with local differential privacy, in: Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 2189–2193.

[39] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J.A. McCann, S.Y. Philip, LoPub, High-dimensional crowdsourced data publication with local differential privacy, IEEE Trans. Inf. Forensics Secur. 13 (9) (2018) 2151–2166.

[40] Q. Ye, H. Hu, X. Meng, H. Zheng, PrivKV: key-value data collection with local differential privacy, in: Proceedings of the 2019 IEEE Symposium on Security and Privacy (S&P), IEEE, 2019, pp. 317–331.

[41] P. Kairouz, S. Oh, P. Viswanath, Extremal mechanisms for local differential privacy, J. Mach. Learn. Res. 17 (1) (2016) 492–542.