2020 16th International Conference on Mobility, Sensing and Networking (MSN) | 978-1-7281-9916-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/MSN50589.2020.00027

An Efficient Data Aggregation Scheme with Local Differential Privacy in Smart Grid

Na Gai*, Kaiping Xue*†¶, Peixuan He†, Bin Zhu*, Jianqing Liu ‡, Debiao He §
* School of Cyber Security, University of Science and Technology of China, Hefei, Anhui 230027, China
† Department of EEIS, University of Science and Technology of China, Hefei, Anhui 230027, China
‡ Department of ECE, University of Alabama in Huntsville, Huntsville, AL 35899, USA
§ School of Cyber Science and Engineering, Wuhan University, Wuhan, Hubei 430072, China
¶Corresponding author, kpxue@ustc.edu.cn

***Abstract*—Smart grid achieves reliable, efficient and flexible grid data processing by integrating traditional power grid with information and communication technology. The control center can evaluate the supply and demand of the power grid through aggregated data of users, and then dynamically adjust the power supply, price of the power, etc. However, since the grid data collected from users may disclose the user's electricity using habits and daily activities, the privacy concern has become a critical issue. 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 the trusted third party. In this paper, we propose a privacy-preserving smart grid data aggregation scheme satisfying local differential privacy (LDP) based on randomized response. Our scheme can achieve efficient and practical estimation of the statistics of power supply and demand while preserving any individual participant's privacy. The performance analysis shows that our scheme is efficient in terms of computation and communication overhead.**

***Index Terms*—Local Differential Privacy, Data Aggregation, Smart Grid, Privacy Preserving**

I. Introduction

Smart grid is considered to be the next generation power grid which provides more intelligent services, such as end-to-end communication and real-time data management, by combining advanced information and communication technology (ICT) [1]. In smart grid, the smart meter installed in each house reports the electricity consumption data to the control center periodically [2, 3]. The control center gathers all the data submitted from smart meters, performs statistical analysis, and then manages the electricity generation, transmission and distribution in the smart grid. Through smart grid, the control center is able to estimate the power consumption of the grid and formulate dynamic pricing strategy.

Despite the promise of smart grid, the reported data from smart meters always contain private information from users. Through these data, it is possible to analyze a user's electricity usage pattern, which causes great threat to his/her privacy [4, 5]. 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 smart grid have been proposed. Homomorphic encryption [3, 6–9] is widely used in data aggregation for preserving data privacy. Homomorphic encryption enables entities to transform operations on plaintext into operations on corresponding ciphertext. Using the homomorphic encryption, especially the semi-homomorphic encryption such as the Paillier cryptosystem which supports addition operation on the ciphertext, the smart meters encrypt the data and send the ciphertext to the aggregator. Then the aggregator integrates the ciphertext gathered from smart meters and decrypts the aggregation result. Homomorphic encryption-based schemes make it possible to preserve single user's data privacy. However, a general problem is that homomorphic encryption brings heavy computation burden to the smart meters, which usually do not have sufficient computing resources. Moreover, given the fact that smart meters submit data periodically and frequently, the homomorphic encryption is not a practical solution for privacy preservation.

Another technique for preserving data privacy in smart grid is to use data masking [10–12]. 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) [13, 14]. Differential privacy [15] has been considered a mathematically rigorous framework for privacy protection and has been adopted in many massive data aggregation and processing scenarios with privacy protection requirements. In the schemes based on DP, data submitted is usually masked by random noise such as Laplacian noise. A common problem in these schemes is that a trusted third party (i.e., a curator) is usually needed for distributing the noise value, but this assumption is not always practical. Although some schemes do not need trusted third party, these distributed approaches usually bring more communication overhead between users.

Recently, local differential privacy (LDP) [16–21] has obtained much attention in both academia and industry. The main idea for LDP is that users perform random perturbations on their data locally. Thus, local differential privacy enables

978-1-7281-9916-0/20/$31.00 ©2020 IEEE
DOI 10.1109/MSN50589.2020.00027

privacy-preserving data collection and aggregation without relying on a trusted third party. Besides, the computation overhead is much lighter compared to the schemes that adopt homomorphic cryptosystem. A typical example of LDP implementation is the RAPPOR [19] developed by Google. RAPPOR enables the Google browser to collect statistical information from end-users while providing strong privacy protection. However, most of LDP schemes focus on the frequency estimation and distribution estimation, and mainly deal with discrete category data.

Considering the needs of data privacy protection and the limited computation resources of smart meters in smart grid, we propose a practical and efficient privacy-preserving data aggregation scheme for demand estimation (in numerical values). By introducing LDP into the data aggregation, the data privacy can be protected without a trusted third party. The computation overhead for each smart meter is acceptable compared to the schemes based on cryptography algorithm. Besides, our scheme has natural support for users' dynamic joining and exiting, the extra cost of which is quite small. The major contributions of our scheme are as follows:

- We propose a lightweight privacy-preserving data aggregation scheme in smart grid based on LDP, in which smart meters can perturb their generated data by randomized response locally without a trusted third party. Meanwhile, our scheme can effectively support users' dynamically joining and exiting without involving much extra overhead.
- To ensure the utility of aggregated data, we design a simple but effective data discretization algorithm based on conditional probability, which can reduce deviation between the aggregation result and the actual data aggregation results. In such a way, our scheme largely increases statistical accuracy of data aggregation results.
- 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.

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 II. Then we describe the problem model including network model and security assumption of our work in the section III. Preliminaries related to our scheme is given in section IV. And details of our LDP-based data aggregation scheme is shown in section V. Then we give the numerical analysis in section VII. Finally, we conclude our work in section VIII.

II. RELATED WORK

A. Privacy-Preserving Data Aggregation in Smart Grid

Data aggregation is a basic service in smart grid, and the privacy protection is one of its primary considerations. To this end, many privacy-preserving data aggregation schemes have been proposed. Homomorphic encryption is one of the most popular methods adopted in privacy-preserving data aggregation schemes, which allows computation on the ciphertext. In [6], Paillier cryptosystem [22] is introduced to construct a privacy-preserving aggregation scheme for secure smart grid. In what follows, many other works [8, 23] based on Paillier cryptosystem have been 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 [24, 25] 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 unpractical to utilize homomorphic encryption to protect data privacy in the smart grid scenario.

To protect the privacy of the data from single user, researchers also proposed some schemes based on data masking [10, 12, 14, 26, 27]. In these schemes, the data submitted by users are masked by a masking value, thus the other entities can not access the real value without knowing the masking value. In [12, 27], 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 [12], 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 [11] that do not depend on the 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, which once again increases the communication overhead.

B. Local Differential Privacy

Local differential privacy [16] has been proposed to provide privacy protection for distributed scenarios where users perturb their data locally and upload without any trusted third party. At present, most of the schemes satisfying LDP are realized by randomized response (RR) [28]. There are also other schemes based on information compression [29] and other disturbance mechanisms to achieve local differential privacy. RR is initially designed to sensitive questions with binary answer "yes" or "no". Users decide to upload the original answer or reverse the answer depending on coin flipping. RR is then easily extended to make statistics on categorical data for frequency estimation. RAPPOR [19] developed by Google encodes data as a Bloom filter and does the randomized response on each bit of the Bloom filter. Wang *et al.* [20] proposed a protocol for finding frequent items in the set-valued LDP setting. Ren

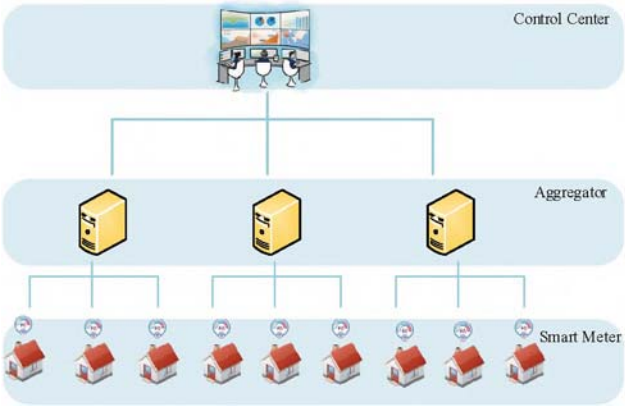

Fig. 1. System Model

et al. [30] focused on the high-dimensional crowdsourced data publication with guarantee of LDP. In [18], discrete distribution estimation based on k-subset mechanism satisfying LDP has been proposed. Most of the works on LDP focus on the 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 smart grid.

III. PROBLEM MODEL

A. Network Model

The network model consists of three kinds of entities, 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 which is installed in 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 power consumption data of a single user. For clarity, we treat the "smart meter" and "user" the same and may use them interchangeably.
- **Gateway (AG):** AG 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.
- **Control Center (CC):** CC connects with all AGs, and collects the data of total power consumption from AGs. Then, CC formulates power dispatching strategy and adjusts electricity price.

B. Security Assumption

In the smart grid, The CC and gateway are managed by electric supply companies, so it can be considered that the CC and AG are honest but curious. They will process the data according to the protocol, but they also interested in the privacy of users' data. We regard the users as honest participant who submit their power consumption data to the gateways, meanwhile they are concerned about their data privacy.

C. Security and Design Goals

Based on the above assumption of each entity, our scheme are proposed to achieve 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.
- *Practicability.* Data submission in smart grid is periodic and frequent. Therefore, efficiency for the data processing and submission becomes important. Since the smart meter does not have plenty of computation capabilities, the computation overhead of each smart meter should be bearable. Also, the communication overhead between users should be as small as possible.
- *No need for 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 impossible to assume a trusted third party. Therefore, our proposed scheme should not rely on a trusted third party.
- *Support dynamic changes of users.* In smart grid, users may join or quit the system, so the aggregation scheme should accommodate the dynamic change of users. More specifically, when some users join or leave, the other users in the system do not need to renegotiate new parameters.

IV. PRELIMINARIES

A. Local Differential Privacy (LDP)

The formal definition of local differential privacy is as follow:

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

$$Pr[\mathcal{M}(X^i) = \widetilde{X}] \leqslant e^{\epsilon} \times Pr[\mathcal{M}(X^j) = \widetilde{X}],$$

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 $\widetilde{X}$ cannot determine whether the input is X^i or X^j.

B. k-Randomized Response (k-RR)

The k-Randomized Response (k-RR) is a 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

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

where ϵ is the privacy budget and $k = |\mathcal{X}|$. 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. References [28, 31] provide more details about k-Randomized Response (k-RR).

V. Scheme Description

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

A. Overview

To estimate the power consumption of the smart grid, the CC needs to get the statistical information of power consumption over a period of time. Therefore, our goal is to design an efficient and practical smart grid data aggregation scheme. The main idea is to discretize data and estimate the total or average power consumption by analyzing data frequency through RR. However, 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. By combining it with randomized response, we design an aggregation protocol satisfying LDP for numerical data.

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

It is noteworthy that the discrete interval division in our scheme needs not to be the same, and it can be decided by the aggregator according to the data analysis demand.

B. System Initialization Phase

In the initialization phase, the gateway first determines the range of data. Since power consumption data that is generated by smart meters are always in a certain range, it is reasonable to assume that the raw data submit by honest users are within the interval $[0, m]$. Then the gateway divides the interval into $[0, s), [s, 2s), ..., [(d-1)s, ds]$, assuming that $d = \lceil \frac{m}{s} \rceil$. For the sake of presentation clarity, here the interval is split evenly into several subintervals. While in practice, the gateway can divide the intervals into arbitary lengths according to the demand. It is noted 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 gateway broadcasts the interval $[0, m]$, the subintervals $[0, s), [s, 2s), ..., [(d-1)s, ds]$ and the privacy budget ϵ to all of the smart meters.

C. Data Submission Phase

For a user u_i with power consumption data x_i which 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 convenience, here we express this subinterval as $[u, v)$.

1) First, user u_i discretizes the real number 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:

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

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

2) Then, user u_i uses the discretized value x_i' to generate the submitted data by 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:

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

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

D. Data Aggregation and Analysis

After receiving y_i from all the 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. Here we denote the frequency of each element $X_j \in X, 0 \leq j \leq |X|$ as $C(X_j)$. And the gateway computes

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

Then the gateway can estimate the total power consumption as

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

After this, the gateways in the system send the total power consumption data to the CC, which can then get statistical information by analyzing them, such as the average and peak of the power consumption.

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

VI. 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 situation of uneven distributed interval.

A. 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 that given in the section IV. Assuming that any two elements x_i, x_j in X, we can have

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

In our scheme, the final submitted data y_i satisfies

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

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

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

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

We can see that the data collected by the control center are perturbed locally and satisfy LDP. Therefore, CC can only aggregate and analyze the data submitted by all smart meters to get 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.

B. 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 $\dfrac{v-x_i}{v-u}$ and gets v with probability $\dfrac{x_i-u}{v-u}$, therefore, the expectation of x_i is

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

It can be seen that, 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 is 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 the actual aggregation result.

C. Support for User Dynamics

In smart grid, there may be the join and exit of smart meters that participate in data aggregation. For each smart meter in the system, it can conduct discretization and perturbation locally with the input: data range, the division of subintervals and the privacy budget, which makes it very convenient for new users to join in the system.

When there exists 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.

D. Situation of Uneven Distributed Interval

In section V, we divide the interval of submitted data equally in order to explain the content of the scheme more clearly. While in practice, the interval can be distributed into unevenly subintervals. The AG can reduce the interval near the average power consumption of users according to the experience. This will not threat users' data privacy, and at the same time, CC can analyze the general power consumption habits of users in the region to a certain extent.

VII. 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 numbers of subintervals s. Then we compare our scheme with two typical data aggregation schemes [10, 14] in smart grid in terms of computation overhead and communication overhead. In [10], the smart meters add noise to the raw data and uploads it to the gateway through homomorphic encryption, and then the gateway obtains the final result through calculation on ciphertext. In [14], the privacy protection of smart meters' data is realized by generating and distributing random numbers for each smart meter in advance.

All 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. In the experiments, if there is no additional statements, 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.

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

A. *Utility Analysis*

We consider different privacy budgets ϵ and numbers of subintervals s to observe the utility of statistical results of our scheme. We randomly generate 1,000 numbers in [0,100] to simulate one thousand user data in data aggregation tasks, of which the correct aggregation result is 48,155. As shown in Fig. 2 and Fig. 3, the circle dotted line represents the actual value of the data aggregation task.

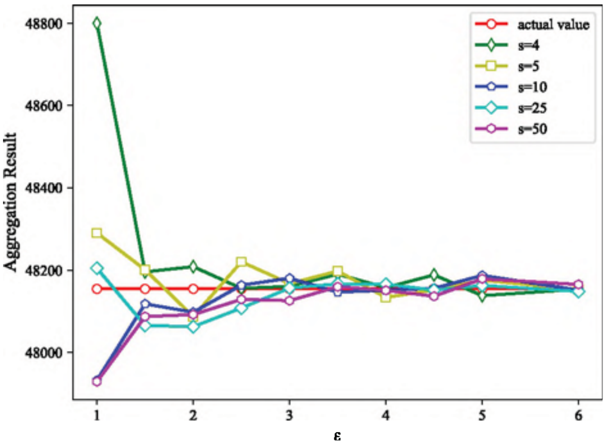

Fig. 2. Statistical result with the relationship of privacy budget and data accuracy

We first evaluate the impact of privacy budgets on the statistical analysis, and Fig. 2 shows the results under different privacy budgets. We can see that with the same number of subintervals s, the larger privacy budgets ϵ is, the closer the statistical results is to the real value. This result is in line with the expectation of theoretical analysis. When the privacy budgets ϵ are small, the probability of the submitted data falling into other intervals is much greater than that of the original interval, which makes the estimated value largely deviate from the real value. Nevertheless 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 choose different privacy budgets according to 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. 3 shows the results with different subinterval numbers s. When $\epsilon = 1$, the results fluctuate as the subinterval number grows. While when the value of privacy budget is reasonable (larger than 1.5 according to aforementioned experiment), the results keep steady with acceptable errors. Thus subinterval number is not the main factor that influence the utility performance.

B. *Computation Overhead*

In this section, we analyse and compare our scheme with Gope's scheme [10] and Bao's scheme [14].

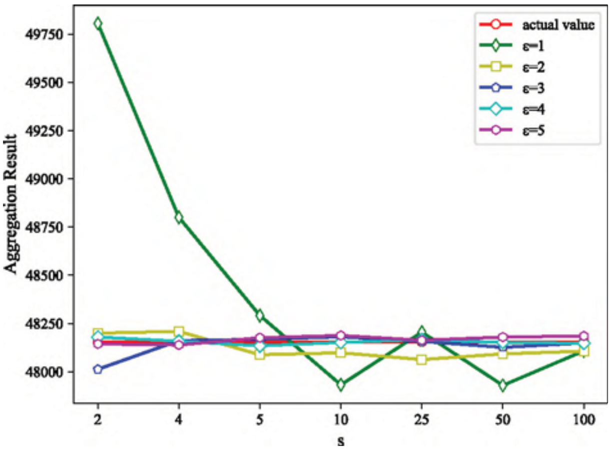

Fig. 3. Statistical result with the relationship of subinterval number and data accuracy

TABLE I
Computation Overhead Comparison

Scheme	Operations	
	Smart Meter	Aggregator
Gope's Scheme [10]	H	$n(SE+H)$
Bao's Scheme [14]	$2*C_e+C_m$	$(n-1)C_m$
Our Scheme	$3ADD+1MUL$	$\lvert X\rvert(4ADD+3MUL)$

* 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.

TABLE I shows the comparison result of computation overhead of smart meters and aggregators (gateways in our scheme). We can see that smart meters and aggregators only need to conduct several times of real number addition and multiplication, which bring little computation overhead. To keep privacy preserved, in Gope's Scheme, smart meters will conduct one hash operation, and aggregators need to conduct n hash operations and symmetric encryptions. And Bao's scheme includes several exponentiation and multiplication over a cyclic group G.

Fig. 4 shows experimental results. Since the overhead of exponentiation and multiplication over G is much larger than other operations, Bao's scheme consumes longest time during aggregation process. On the contrary, benefit from the lightweight real number addition and multiplication, our scheme has the best performance, with 0.0225 ms time cost at smart meters and 0.0454 ms time cost at aggregators.

Fig. 5 shows the total computation overhead of the aggregators with varying numbers of smart meters participating in the data aggregation task. When the number of smart meters participating in the task is not large, the computation cost of the three schemes are all acceptable. However, when the number of participating meters increases, the computation overhead of Gope's scheme and Bao's scheme will grow approximately linearly. While in our scheme, the computation

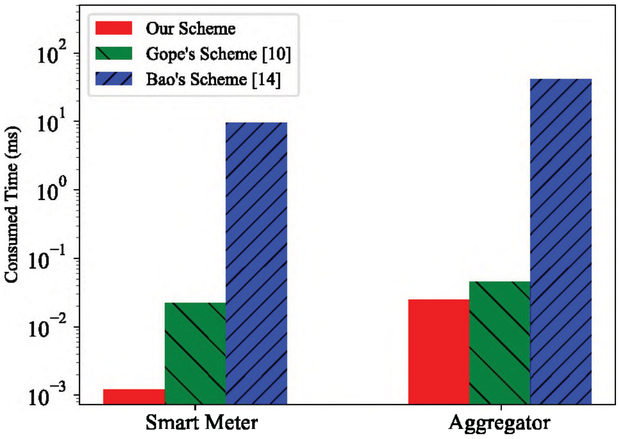

Fig. 4. Computation overhead of different entities

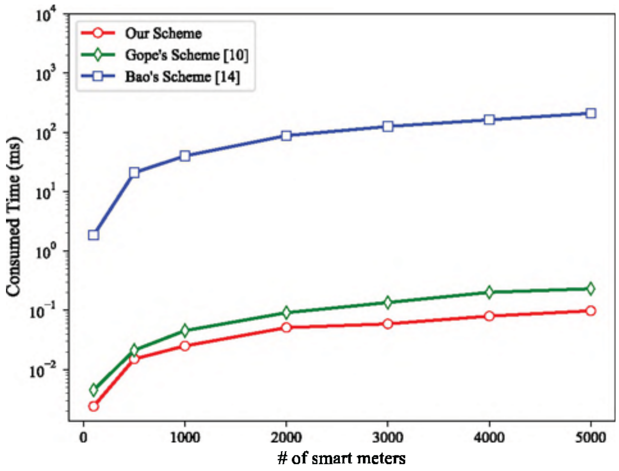

Fig. 5. Computation overhead of the aggregators with varying numbers of smart meters

overhead of the aggregators remains stable, of which the reason is that the aggregators only has to count the number of each discrete value and calculate the final result. Therefore, the computation overhead is much more 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. In addition, the tasks of data aggregation is often very frequent in smart grid, which requires the computation overhead of every single data aggregation task to be as small as possible. Through the analysis of computing overhead, we can conclude that our scheme is more efficient and practical than other schemes in the real-world smart grid system.

C. Communication Overhead

In this section, we analyse communication overhead of our scheme and compared two schemes. The total communication overhead of a single aggregation task with different numbers of participating smart meters is shown in Fig. 6. Due to the large range of communication overhead, we use logarithmic coordinates to show the experimental results.

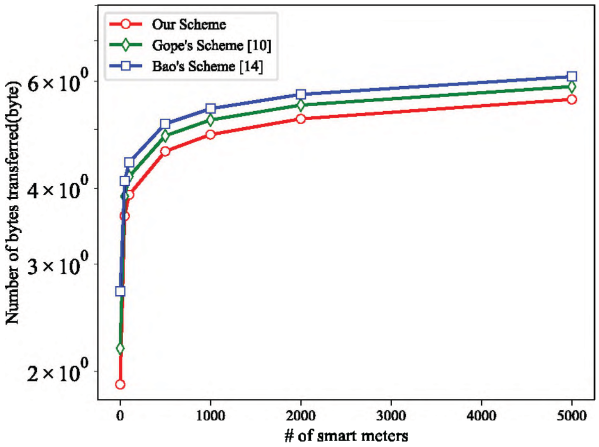

Fig. 6. Communication overhead with varying numbers of smart meters

For fairness consideration, in Gope's scheme, we only consider the communication cost in the data aggregation part of the scheme for comparison. As shown in Fig. 6, the communication overhead of these three schemes increases with the growing number of smart meters. Since there is no encryption in our scheme, only the division of the interval, the privacy budget, perturbed data and aggregation results are needed to be transmitted during the aggregation process. Considering the large volume of ciphertext, our scheme has the least communication overhead. While in Gope's scheme, besides necessary perturbed data and aggregation results, it needs to transmit extra 2 hash values and one ciphertext. Thus communication overhead of Gope's scheme is larger than ours. Besides, 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.

VIII. Conclusion

In this paper, we proposed a privacy-preserving data aggregation scheme for the smart grid. Considering the limited computation ability of smart meters, we reduced the computation burden of smart meters that participate in data aggregation tasks. By designing a special data discretization algorithm and randomized response mechanism, the scheme achieves the privacy-preserving smart grid data aggregation which satisfies the LDP. Unlike existing schemes based on masking values, our scheme is able to run normally without a trusted third party. Since users need not to negotiate with each other for the masking values, 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 literatures.

Acknowledgment

This work is supported in part by the National Key Research and Development Program of China under Grant No. 2016YFB0800301, the National Natural Science Foundation of China under Grant No. 61972371, Youth Innovation Promotion Association Chinese Academy of Sciences (CAS) under Grant No. 2016394, and and the Fundamental Research Funds for the Central Universities.

References

[1] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. P. Hancke, "Smart grid technologies: Communication technologies and standards," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 4, pp. 529–539, 2011.

[2] A. Abdallah and X. S. Shen, "A lightweight lattice-based homomorphic privacy-preserving data aggregation scheme for smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 396–405, 2016.

[3] Y. Liu, W. Guo, C.-I. Fan, L. Chang, and C. Cheng, "A practical privacy-preserving data aggregation (3PDA) scheme for smart grid," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1767–1774, 2018.

[4] H. Khurana, M. Hadley, N. Lu, and D. A. Frincke, "Smart-grid security issues," *IEEE Security & Privacy*, vol. 8, no. 1, pp. 81–85, 2010.

[5] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security & Privacy*, vol. 7, no. 3, pp. 75–77, 2009.

[6] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.

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

[8] T. W. Chim, S.-M. Yiu, V. O. Li, L. C. Hui, and J. Zhong, "PRGA: Privacy-preserving recording & gateway-assisted authentication of power usage information for smart grid," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 1, pp. 85–97, 2014.

[9] L. Chen, R. Lu, and Z. Cao, "PDAFT: A privacy-preserving data aggregation scheme with fault tolerance for smart grid communications," *Peer-to-peer Networking and Applications*, vol. 8, no. 6, pp. 1122–1132, 2015.

[10] P. Gope and B. Sikdar, "An efficient data aggregation scheme for privacy-friendly dynamic pricing-based billing and demand-response management in smart grids," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3126–3135, 2018.

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

[12] W. Jia, H. Zhu, Z. Cao, X. Dong, and C. Xiao, "Human-factor-aware privacy-preserving aggregation in smart grid," *IEEE Systems Journal*, vol. 8, no. 2, pp. 598–607, 2013.

[13] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer, 2006.

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

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, pp. 265–284, Springer, 2006.

[16] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy, data processing inequalities, and statistical minimax rates," pp. 1592–1597, 2013.

[17] C. Xu, J. Ren, D. Zhang, and Y. Zhang, "Distilling at the edge: A local differential privacy obfuscation framework for iot data analytics," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 20–25, 2018.

[18] S. Wang, L. Huang, Y. Nie, X. Zhang, P. Wang, H. Xu, and W. Yang, "Local differential private data aggregation for discrete distribution estimation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 9, pp. 2046–2059, 2019.

[19] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, ACM, 2014.

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

[21] J. Liu, C. Zhang, and Y. Fang, "EPIC: A differential privacy framework to defend smart homes against internet traffic analysis," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1206–1217, 2018.

[22] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 223–238, Springer, 1999.

[23] S. Li, K. Xue, Q. Yang, and P. Hong, "PPMA: Privacy-preserving multisubset data aggregation in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 462–471, 2017.

[24] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-DNF formulas on ciphertexts," in *Proceedings of 2005 Theory of cryptography conference (TCC)*, pp. 325–341, Springer, 2005.

[25] H. Bao and R. Lu, "A new differentially private data aggregation with fault tolerance for smart grid communications," *IEEE Internet of Things Journal*, vol. 2, no. 3, pp. 248–258, 2015.

[26] C. Castelluccia, A. C. Chan, E. Mykletun, and G. Tsudik, "Efficient and provably secure aggregation of encrypted data in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 5, no. 3, p. 20, 2009.

[27] L. Lyu, K. Nandakumar, B. Rubinstein, J. Jin, J. Bedo, and M. Palaniswami, "PPFA: privacy preserving fog-enabled aggregation in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3733–3744, 2018.

[28] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

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

[30] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and S. Y. Philip, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.

[31] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in neural information processing systems*, pp. 2879–2887, NIPS, 2014.