sensors

MDPI

Article

A Lightweight and Privacy-Friendly Data Aggregation Scheme against Abnormal Data

Jianhong Zhang [1,2,*] and Haoting Han [1]

[1] School of Information Sciences and Technology, North China University of Technology, Beijing 100043, China; ncuthht105@mail.ncut.edu.cn
[2] Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
* Correspondence: zjhncut@163.com

Abstract: Abnormal electricity data, caused by electricity theft or meter failure, leads to the inaccuracy of aggregation results. These inaccurate results not only harm the interests of users but also affect the decision-making of the power system. However, the existing data aggregation schemes do not consider the impact of abnormal data. How to filter out abnormal data is a challenge. To solve this problem, in this study, we propose a lightweight and privacy-friendly data aggregation scheme against abnormal data, in which the valid data can correctly be aggregated but abnormal data will be filtered out during the aggregation process. This is more suitable for resource-limited smart meters, due to the adoption of lightweight matrix encryption. The automatic filtering of abnormal data without additional processes and the detection of abnormal data sources are where our protocol outperforms other schemes. Finally, a detailed security analysis shows that the proposed scheme can protect the privacy of users' data. In addition, the results of extensive simulations demonstrate that the additional computation cost to filter the abnormal data is within the acceptable range, which shows that our proposed scheme is still very effective.

Keywords: data aggregation; abnormal data; source; matrix encryption; lightweight

Citation: Zhang, J.; Han, H. A Lightweight and Privacy-Friendly Data Aggregation Scheme against Abnormal Data. *Sensors* **2022**, *22*, 1452. https://doi.org/10.3390/s22041452

Academic Editors: Shaoen Wu, Jinbo Xiong, Periklis Chatzimisios and Mahmoud Daneshmand

Received: 11 January 2022
Accepted: 10 February 2022
Published: 14 February 2022

1. Introduction

With the application of electricity in our daily life becoming increasingly extensive, more factors need to be considered in the production decisions of the cloud server [1,2], such as how to maintain a balance between supply and demand when electricity usage changes dramatically [3]. Thus, it is critical to obtain the electricity usage data of all users. In addition the smart grid, as a key infrastructure, adds upstream information feedback based on the traditional grid, which can help us collect the electricity usage data of users in various regions [4,5]. The prominent advantage of smart meters is to make sure that electricity supply matches the demand of users within a short period, which is of great significance for the rational distribution of power resources and the reduction of economic losses [6,7]. To obtain the real-time electricity demand of users, their electricity usage data should be measured, aggregated, and analyzed through advanced metering infrastructure [8,9].

However, it is a noteworthy problem of the smart grid that the abnormal electricity data, caused by electricity theft or meter failure, can lead to inaccurate aggregation results. This not only harms the personal interests of users, but also interferes with the production decisions of the cloud center. To the best of our knowledge, none of the existing schemes consider the impact of abnormal data. In the extant schemes, the aggregation center is responsible for aggregating all the reported electricity usage data of smart meters but cannot detect whether the reported data is abnormal, let alone find the source of the abnormal data.

Therefore, it is an important challenge to filter out the abnormal data and find the source of the abnormal data when the data is encrypted. To address this issue, we propose a

lightweight and privacy-friendly data aggregation scheme against abnormal data, in which the valid data is correctly aggregated, but the abnormal data is automatically filtered out during the aggregation process. Notably, the filtration of the abnormal data does not need additional procedures, which is the highlight of this work. Besides, compared with other methods in other schemes, the encryption method used in our scheme is more suitable for smart meters with limited computing capacity. Specifically, the main contributions of this paper are summarized as follows:

- We propose a lightweight and privacy-friendly data aggregation scheme against abnormal data by using lightweight matrix encryption. It is suitable for smart meters with limited computing power, since no time-consuming computation operators are involved.
- Abnormal data can automatically be filtered out without additional procedures. In addition, the source of the abnormal data can also be found out in this process. Thereby, accurate aggregation results can be obtained through the proposed scheme, and abnormal meters can also be identified for maintenance, even if the data is encrypted.
- Finally, a detailed security analysis is provided to prove that our scheme can fully ensure the privacy and security of users' data. Experiments and performance evaluations demonstrate that our scheme has a low computation cost and high practicality.

The rest of the paper is outlined as follows. In Section 2, some related works are provided. The preliminary is provided in Section 3. Section 4 illustrates the system model and adversary model. We propose the details of our scheme in Section 5, followed by the security analysis of our scheme in Section 6. A performance analysis is conducted in Section 7. Finally, the conclusion of our scheme is summarized in Section 8.

2. Related Work

There exist extensive data aggregation schemes on the topic of protecting users' privacy in smart grids [10–23]. Homomorphic encryption has been applied in several works to achieve privacy-preserving data aggregation [10–19]. Shen et al. [10] proposed a Paillier-based data aggregation scheme against malicious data mining attacks, which can prevent the adversary from inferring a target user's electricity usage data and obtain accurate aggregated results of electricity usage data. Xue et al. [11] proposed a privacy-preserving service-outsourcing scheme for a real-time pricing demand response in a smart grid, which solves the privacy issues by modifying the Paillier cryptosystem to hold two different decryption keys and achieves the flexible enrollment and revocation of smart meters. In addition, Saleem et al. [12] proposed a scheme to resist the malfunctioning of smart meters for data aggregation based on a modified Paillier cryptosystem. Their system can resist false data injection attacks by filtering out the inserted values from external attackers. For achieving secure data aggregation, the ElGamal-based algorithm has been taken into account [13,14]. Liu et al. [14] proposed a lifted elliptic ElGamal-based privacy-preserving data aggregation scheme, in which the trusted third party is removed and the users, with some measure of trust, construct a virtual aggregation area to mask the single user's data against the denial of service attack. In order to resist quantum attacks and improve the efficiency of the algorithm, the lattice-based homomorphic approach has been applied to achieve secure data aggregation for smart grids [15,16]. Abdallah et al. [16] proposed a lattice-based privacy-preserving data aggregation scheme for a smart grid, which can further reduce the computation burden for smart appliances, because it depends on simple arithmetic operations. In [17], a privacy-friendly data aggregation scheme is proposed by Vahedi et al. They use elliptic curve digital signature algorithms (ECDSA) in smart grids to protect users' privacy from the grid operators. Besides, to meet the higher data analysis requirements of the cloud server, multidimensional data is aggregated in some schemes [18–20]. Although the schemes based on homomorphic encryption can obtain accurate aggregation results, a heavy computational and communication burden will also be imposed on smart meters with limited computing power.

As another major encryption technology, masking-value-based schemes also have been proposed to achieve secure and efficient data aggregation in smart grids. As for masking-based data aggregation schemes [21–24], Gope et al. [21] first proposed a lightweight and privacy-friendly masking-based spatial data aggregation scheme for secure forecasting of power demands in smart grids. Their scheme only uses lightweight cryptographic primitives, such as exclusive OR operations and hash functions, thus it has a significantly lower computational cost as compared with other approaches. The LCEDA scheme proposed by Su et al. [22] achieves an efficient update of masking the value share to ensure forward security of individual data, dynamic enrollment, and revocation of smart meters. Moreover, Huang et al. [23] propose a lightweight and fault-tolerable data aggregation scheme that can determine the smart meters which fail to upload data on time with the idea of flag bit, and correct aggregation results can be obtained even if the data is not reported by the smart meters. However, the existing masking-based aggregation schemes cannot screen abnormal electricity consumption data either.

In addition, accurate aggregation results can be obtained by utilizing zero-knowledge proof [24], but heavy communication and the computational burden will also be imposed on the smart meters with limited computing power. Thus, the solution using zero-knowledge proof is not practical.

Therefore, we propose a lightweight and privacy-friendly data aggregation scheme against abnormal data by using matrix encryption, which can effectively filter abnormal data and find out the source of abnormal data. To more intuitively show the advantages of the proposed scheme compared with other schemes, the security feature comparisons are shown in Table 1.

Table 1. Security feature comparisons.

Scheme	Data Confidentiality	Resistance to Middle-Man Attacks	Filtering of Abnormal Data	Tracing the Source of Abnormal Data	Computational Cost
[10–12]	✓	✓	×	×	*High*
[13,14]	✓	✓	×	×	*High*
[15,16]	✓	✓	×	×	*Low*
[17]	✓	✓	×	×	*High*
[18–20]	✓	✓	×	×	*High*
[21,22]	✓	✓	×	×	*Low*
[23]	✓	✓	×	×	*Low*
[24]	✓	✓	✓	✓	*High*
The proposed protocol	✓	✓	✓	✓	*Low*

3. Preliminaries

In this section, the preliminaries of the proposed scheme are presented, in which we describe the basic idea of filtering abnormal data.

Filtering abnormal data: Suppose that $\boldsymbol{a}$ is the data to be determined and $\boldsymbol{b}$ is the upper limit of the normal value, and they are in the range of $[0, N^2 - 1]$. Then, whether the data $\boldsymbol{a}$ is abnormal can be determined as follows [25]:

1. Construct an $N \times N$ matrix containing all possible values in $[0, N^2 - 1]$, as shown in Figure 1. Each value has a row coordinate and a column coordinate in this matrix. The value in the matrix can be represented by $iN + j$, the corresponding row coordinate and the column coordinate of this value are $(i + 1)$ and $(j + 1)$, respectively. Based on these values, $\boldsymbol{a}$ and $\boldsymbol{b}$ can be represented as two-dimensional coordinates (i_a, j_a) and (i_b, j_b), where i_a, j_a is the row and column coordinate of $\boldsymbol{a}$, and i_b, j_b is the row and column coordinate of $\boldsymbol{b}$. (i_a, j_a) and (i_b, j_b) can be computed from the following formulae:

$$i_a = \left\lceil \frac{a}{N} \right\rceil + 1; j_a = a \bmod N + 1. \quad (1)$$

$$i_b = \left[\frac{b}{N}\right] + 1; j_b = b \bmod N + 1. \tag{2}$$

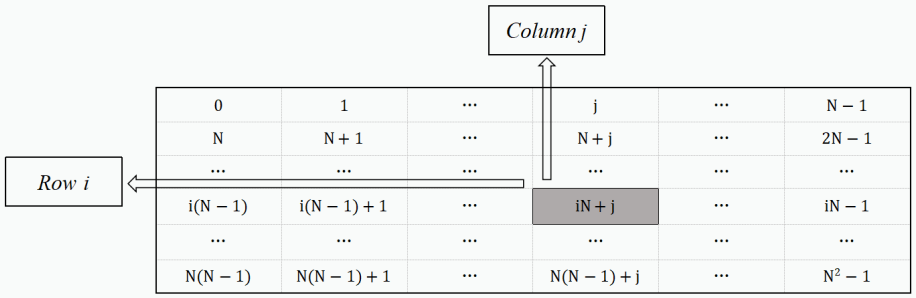

Figure 1. Representation of the constructed matrix.

2. Based on (i_a, j_a), we can construct three N-dimensional column vectors for $\boldsymbol{a}$ as

$$\tilde{a} = \begin{bmatrix} \mathbf{0}^{i_a} \\ \mathbf{1}^{N-i_a} \end{bmatrix}, \bar{a} = \mathbf{e}_{i_a}, \hat{a} = \begin{bmatrix} \mathbf{0}^{j_a-1} \\ \mathbf{1}^{N-j_a+1} \end{bmatrix}, \tag{3}$$

where $\mathbf{0}^{i_a}$ denotes an i_a-dimensional zero vector, $\mathbf{1}^{N-i_a}$ denotes an 1^{N-i_a}-dimensional vector, and all elements are 1; $\mathbf{e}_{i_a}$ denotes an N-dimensional unit vector, and the i_a-th element is 1. In this way, we can obtain the following transformation relation:

$$\begin{cases} a \le b \Leftrightarrow \tilde{a}[i_b] + \bar{a}[i_b] * \hat{a}[i_b] = 1 \\ a > b \Leftrightarrow \tilde{a}[i_b] + \bar{a}[i_b] * \hat{a}[i_b] = 0 \end{cases} \tag{4}$$

3. Construct $X = [\tilde{a}^T \bar{a}^T]$, $X' = [1\ \hat{a}^T]$ and a $2N \times (N+1)$ matrix Q satisfying

$$Q[i_b, 1] = Q[N + i_b, j_b + 1] = 1, \tag{5}$$

and the other elements in Q are 0. We have

$$\tilde{a}[i_b] + \bar{a}[i_b] * \hat{a}[i_b] = [\tilde{a}^T \bar{a}^T] Q \begin{bmatrix} 1 \\ \hat{a} \end{bmatrix} = XQX'^T. \tag{6}$$

So we have the conclusion that

$$\begin{cases} a \le b \Leftrightarrow \tilde{a}[i_b] + \bar{a}[i_b] * \hat{a}[i_b] = 1 \Leftrightarrow XQX'^T = 1 \\ a > b \Leftrightarrow \tilde{a}[i_b] + \bar{a}[i_b] * \hat{a}[i_b] = 0 \Leftrightarrow XQX'^T = 0. \end{cases} \tag{7}$$

As we describe above, the judgment on whether data $\boldsymbol{a}$ is abnormal can be transformed to the equality test of $XQX'^T = 1$ or 0. To be specific, if $XQX'^T = 1$, it is equivalent to the fact that a is less than or equal to b, where b is the upper limit of the normal value we set. Therefore it means that the data a is normal. The opposite is also true.

To help readers better understand the principle, we add a numerical example here. Supposed that $N = 10, a = 55, b = 60$. We have $i_a = \left[\frac{55}{10}\right] + 1 = 6; j_a = 55 \bmod 10 + 1 = 6$, and $i_b = \left[\frac{60}{10}\right] + 1 = 7; j_b = 60 \bmod 10 + 1 = 1$. Therefore, $\tilde{a} = [0000001111]^{\mathsf{T}}$, $\bar{a} = [0000010000]^{\mathsf{T}}$, and $\hat{a} = [0000011111]$. Further, we can obtain $\tilde{a}[7] + \bar{a}[7] * \hat{a}[7] = 1$. Next we construct the matrix $X, X' and Q$ as described above. Finally, we can obtain that $XQX'^T = 1$, which is also equivalent to $a \le b$.

4. System Model

In this section, we will introduce the system model and the adversary model of the proposed scheme.

4.1. System Model

The system model of our scheme is shown in Figure 2, which consists of three entities: smart meters(*SM*), the aggregation center (*AC*), and the cloud server(*CS*).

- Smart Meters (*SM*): Smart meters are intelligent devices installed at users' premises with limited computing resources. Each smart meter encrypts the electricity usage data and reports it to the aggregation center.
- Aggregation Center (*AC*): The aggregation center has sufficient computing power to collect and aggregate the electricity usage data reported by the smart meters.
- Cloud Server (*CS*): The cloud server receives and analyzes the aggregated results sent by the *AC*, thus making appropriate production decisions and reasonable electricity distribution.

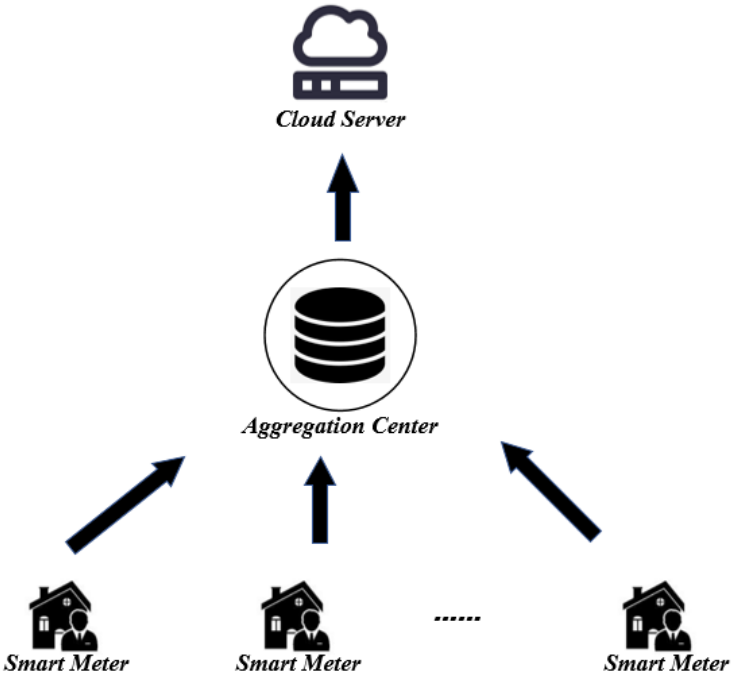

Figure 2. System model.

4.2. Adversary Model

In this scheme, we assume that:

- Users may not only try to steal electricity by compromising smart meters, but also be interested in the privacy of other users' electricity usage data. In addition, there may be cases where the meter fails and reports abnormal electricity consumption data.
- *AC* and *CS* are semi-honest. This means that the two entities will honestly execute the proposed protocol and do not tamper with the computational results, but they may attempt to learn individual electricity usage data as much as possible. Besides, *AC* and *CS* will not collude with each other.
- Any probabilistic polynomial-time adversary can intercept the channels between *SM*s and *AC* and the channels between *AC* and *CS* to obtain the reported data.

Other security issues are beyond the scope of our scheme.

4.3. Security Goals and Functionality

On the basis of the system model and adversary model above, our system should satisfy the following security goals and functionality requirements.

- **Data privacy:** Because the data reported by the electricity meters is closely linked to the users' daily habits and household situations, the proposed scheme should ensure that the privacy of users' electricity usage data is not compromised by curious internal entities, as well as by external attackers.
- **Filter abnormal data:** In order to prevent the abnormal electricity usage data reported by the electricity meters from affecting the accuracy of the aggregation results, the abnormal electricity usage data should be filtered out during the aggregation process.
- **Trace abnormal source:** The proposed scheme should track the source of abnormal data to further repair and maintain abnormal meters.

5. The Proposed Scheme

Our scheme is mainly composed of five stages: system initialization, registration, data encryption, aggregation and filtering, and decryption. In addition, the work flow of our scheme is presented in Figure 3.

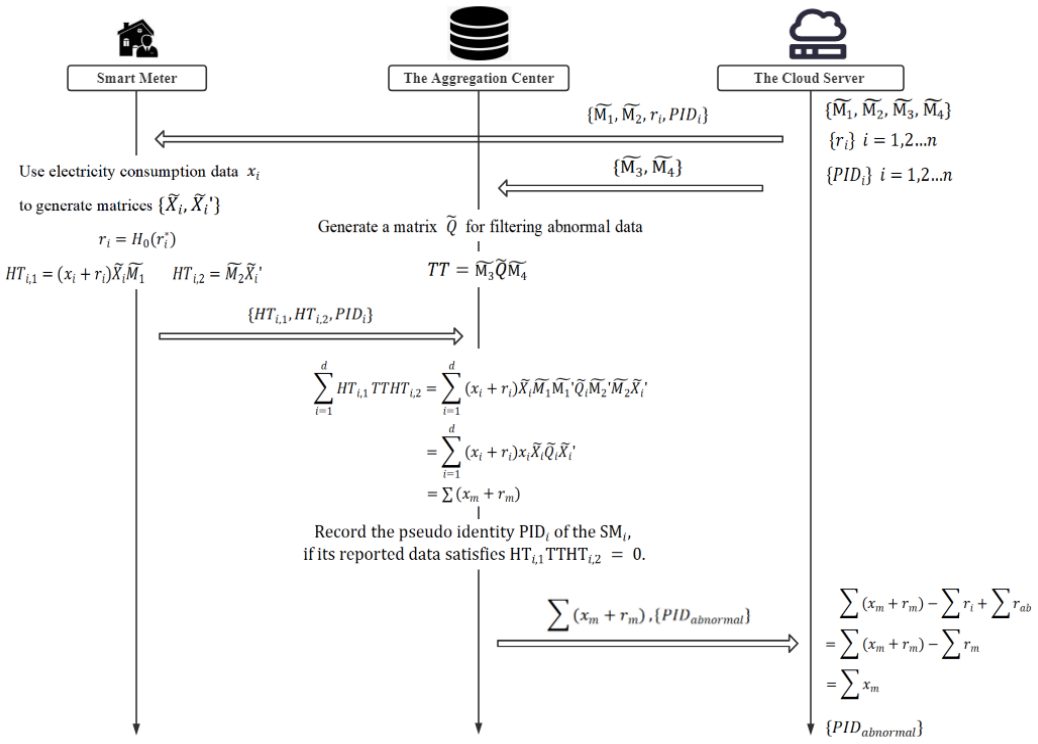

Figure 3. The work flow of our scheme.

5.1. System Initialization

At this stage, the cloud server generates two random non-singular matrices, $\tilde{M}_1 \in R^{(2N+4)\times(2N+4)}$ and $\tilde{M}_2 \in R^{(N+5)\times(N+5)}$, and computes their inverse matrices, $\tilde{M}_3$, $\tilde{M}_4$. After that, the public parameters of the system can be denoted as $\tilde{M}_1$, $\tilde{M}_2$, $\tilde{M}_3$, $\tilde{M}_4$, and H_0. Here, the symbol H_0 is a collision-resistant one-way hash function.

5.2. Registration

When the smart meter SM_i registers with the cloud server, the cloud server generates a random number r_i and a pseudo-identity PID_i for it. Then, the cloud server sends $\{PID_i, r_i\}$ to it over a secure channel.

5.3. User: Data Encryption

(1) For electricity usage data x_i, the smart meter SM_i generates two random numbers, $\mu_{x,i}$ and $\mu'_{x,i}$, and constructs the following matrices $\left\{\tilde{X}_i, \tilde{X}'_i\right\}$ as

$$\tilde{x_i} = \begin{bmatrix} \mathbf{0}^{i_x} \\ \mathbf{1}^{N-i_x} \end{bmatrix}, \bar{x_i} = \mathbf{e}_{i_x}, \hat{x_i} = \begin{bmatrix} \mathbf{0}^{j_x-1} \\ \mathbf{1}^{N-j_x+1} \end{bmatrix}, \tag{8}$$

where $\tilde{x_i}$, $\bar{x_i}$, and $\hat{x_i}$ are constructed as in Section 3, i.e.:

$$\mathrm{X_i} = [\tilde{x}_i^T \ \bar{x}_i^T], \mathrm{X'_i} = [1 \ \hat{x}_i^T] \tag{9}$$

$$\tilde{\mathrm{X}}_\mathrm{i} = [(x_i + r_i)X_i \ R_{x,i} \ 1 \ 0] \tag{10}$$

$$\tilde{\mathrm{X}}'_\mathrm{i} = \begin{bmatrix} X_i'^T \\ R'_{x,i} \\ 0 \\ 1 \end{bmatrix}, \tag{11}$$

where $R_{x,i} = [\mu_{x,i} \ \mu_{x,i}]$, $R'_{x,i} = [\mu'_{x,i} \ \mu'_{x,i}]$.

(2) The smart meter SM_i encrypts $\left\{\tilde{X}_i, \tilde{X}'_i\right\}$ into the ciphertext $\{HT_{i,1}, HT_{i,2}\}$ as follows:

$$HT_{i,1} = \tilde{X}_i \tilde{M}_1, \ HT_{i,2} = \tilde{M}_2 \tilde{X}'_i. \tag{12}$$

(3) Finally, the smart meter SM_i reports the ciphertext $\{HT_{i,1}, HT_{i,2}, PID_i\}$ to the aggregation center.

5.4. The Aggregation Center: Aggregation and Filtering

(1) The aggregation center generates the matrix $\tilde{Q}$ according to the upper limit of normal data, $\boldsymbol{q}$:

$$\tilde{Q} = \begin{bmatrix} Q & 0 & 0 & 0 \\ 0 & R_Q & 0 & 0 \\ 0 & 0 & \mu_{Q,1} & 0 \\ 0 & 0 & 0 & \mu_{Q,2} \end{bmatrix}, \tag{13}$$

where $\mu_{Q,1}$ and $\mu_{Q,2}$ are random numbers, and Q is a $2N \times (N+1)$ matrix constructed as in Section 3:

$$\mathrm{R_Q} = \begin{bmatrix} r_{Q,1} & r_{Q,2} \\ r_{Q,3} & -(r_{Q,1} + r_{Q,2} + r_{Q,3}) \end{bmatrix}, \tag{14}$$

where $r_{Q,1}$, $r_{Q,1}$, and $r_{Q,1}$ are random numbers.

Then, the aggregation center constructs matrix TT according to the matrix $\tilde{Q}$ and the matrix $\left\{\tilde{M}_3, \tilde{M}_4\right\}$ as in the following equation:

$$TT = \tilde{M}_3 \tilde{Q} \tilde{M}_4. \tag{15}$$

(2) The aggregation center aggregates the reported data to obtain the aggregation result R' according to the following equation:

$$\begin{aligned} R' &= \sum_{i=1}^{n} HT_{i,1}TTHT_{i,2} \\ &= \sum_{i=1}^{n} (x_i + r_i)\tilde{X}_i\tilde{M}_1\tilde{M}_3\tilde{Q}_i\tilde{M}_4\tilde{M}_2\tilde{X}_i' \\ &= \sum_{i=1}^{n} (x_i + r_i)\tilde{X}_i\tilde{Q}_i\tilde{X}_i' \\ &= \sum_{i=1}^{n} (x_i + r_i)(XQX'^T + R'_{x,i}R_QR_{x,i}) \\ &= \sum_{i=1}^{n} (x_i + r_i)XQX'^T. \end{aligned} \tag{16}$$

For abnormal data, the result of XQX'^T is 0, therefore the result of the formula $HT_{i,1}TTHT_{i,2}$ is 0. While, for normal data, $XQX'^T = 1$, the result of the formula $HT_{i,1}TTHT_{i,2}$ is still $(x_i + r_i)$. In this way, the abnormal data is automatically filtered in the process of aggregation, that is, the aggregation result R' is $\sum(x_m + r_m)$, where x_m represents the normal electricity usage data, and r_m represents its corresponding masking value. Besides, if reported data are judged to be abnormal, the aggregation center will record their source, PID_{ab}, and send it to the cloud server.

$$(x_i + r_i)XQX'^T = \begin{cases} (x_i + r_i), if\ data\ \ x_i\ \ is\ normal \\ 0, otherwise \end{cases} \tag{17}$$

(3) Lastly, the aggregation center sends the aggregated result, $R' = \sum(x_m + r_m)$, and the pseudo identities, $\{PID_{ab}\}$, of the abnormal smart meters to the cloud server.

5.5. The Cloud Server: Decryption

After receiving the aggregated result, $R' = \sum(x_m + r_m)$, and the pseudo identities, $\{PID_{ab}\}$, of the abnormal smart meters from the aggregation center, the cloud server decrypts the data to obtain the real aggregated result R as in the following equation:

$$\begin{aligned} R &= \sum(x_m + r_m) - (\sum r_m) \\ &= \sum(x_m + r_m) - (\sum r_i - \sum r_{ab}) \\ &= \sum x_m, \end{aligned} \tag{18}$$

where r_{ab} represents the masking value corresponding to the smart meter which reports abnormal data.

Therefore, the cloud server can obtain the accurate aggregated result R that does not include abnormal data and the pseudo identities PID_{ab} of abnormal meters, so that it can make appropriate production decisions and check for abnormal smart meters.

As the range of electricity usage data expands, the constructed matrix will become larger, which greatly increases the communication cost. For example, the bit length of the report data will be at least 1000 bits when the electricity usage data reaches 1000.

To solve this problem, a mapping function $f : S \rightarrow S^*$ is proposed to map the original data to a smaller set, where S and S^* are the original data set and the mapped set, respectively. For any $x_i \in S$, there exists a unique $x_i^* \in S^*$ corresponding to it and $x_i^* = [x_i/b]$, where b is determined by the filtering accuracy. By sacrificing some accuracy within an acceptable range, communication overheads can be greatly reduced.

6. Security Analysis

In this section, we present the security proof of the proposed scheme to solve the problem of adversarial models.

Theorem 1. *(Resistant to the middle-man attack) The proposed scheme can ensure that the privacy of users' data is not compromised by the external adversaries.*

Proof. The confidentiality of users' electricity data $x_i (i = 1, 2, \ldots, n)$ and the aggregation result $\sum x_m$ will be proved below.

If the PPT adversary tries to obtain x_i from $\{HT_{i,1},\ HT_{i,2}\}$, (s)he must know r_i since $HT_{i,1} = (x_i + r_i)\tilde{X}_i\tilde{M}_1$ and $HT_{i,2} = \tilde{M}_2\tilde{X}_i'$. However, r_i is a random number only available to registered users and the cloud server. Consequently, the external adversaries cannot infer the individual electricity data x_i from $\{HT_{i,1},\ HT_{i,2}\}$.

If the external adversary tries to derive $\sum x_m$ from R, (s)he needs to know the sum of random numbers $\sum r_m$ since $R = \sum (x_m + r_m)$. However, $\sum r_m$ is only available to the cloud server. Thus, adversaries cannot infer the normal total electricity usage data $\sum x_m$.

To sum up, any adversary cannot recover individual electricity usage data x_i or total electricity usage data $\sum x_m$ that excludes abnormal data. □

Theorem 2. *Our proposed scheme can achieve the privacy of data transmitted by a smart meter.*

Proof. In our scheme, the attackers of data privacy can be divided into two categories: internal attackers and external attackers. For external attacks, they can be resisted, since an encryption algorithm is adopted in our scheme. For internal attackers, we discuss it in the following three cases. □

1. When the internal attacker is the aggregation center, although it can obtain the encrypted users' electricity usage data, it cannot gain the users' real electricity usage data. Specifically, the aggregation center can get $\{HT_{i,1}, HT_{i,2}\}$ reported by smart meters, where $HT_{i,1} = (x_i + r_i)\tilde{X}_i\tilde{M}_1$, and $HT_{i,2} = \tilde{M}_2\tilde{X}_i'$. If the aggregation center tries to recover x_i from $\{HT_{i,1}, HT_{i,2}\}$, it must know r_i. However, r_i is only available to the user i and the cloud server. Therefore, the proposed scheme can resist privacy attacks on the transmitted data from the aggregation center.
2. When the internal attacker is the cloud server, although it can obtain the aggregated result of normal electricity usage data, it cannot gain the electricity usage data of a single user. Concretely, the cloud server can only obtain $\sum (x_m + r_m)$ from the aggregation center, that is, it can only obtain the aggregated result of normal electricity usage data $\sum x_m$, which is computed by $\sum (x_m + r_m) - \sum r_m$. Therefore, the proposed scheme can resist privacy attacks on the transmitted data from the cloud server.
3. When the internal attacker is a valid smart meter. Although it can intercept electricity usage data reported by other smart meters, it cannot obtain that the corresponding user's true electricity usage information x_i, because the masking value r_i is known only to the corresponding user and the cloud server. Hence, any smart meter cannot recover the electricity usage data of other smart meters.

To sum up, our scheme can achieve data privacy.

Theorem 3. *It is infeasible to learn users' electricity usage data information according to the reported data in different rounds.*

Proof. In each round of data aggregation, the smart meter SM_i updates the masking value r_i as $r_i' = H_0(r_i)$. Even if the adversary gets the reported data in two different rounds, $(x_i + r_i)$ and $(x_i' + r_i')$, (s)he can only obtain $(x_i' + r_i') - (x_i + r_i)$, which does not reveal the changes in electricity usage data in the two aggregation rounds. Therefore, it is still infeasible to obtain information related to users' electricity usage data according to the reported data in different rounds. □

Finally, we compare the security features of our proposed approach with homomorphic encryption schemes [10] and masking-value-based schemes [22,26–28]. As shown in Table 2, our scheme has the most comprehensive security functions and features.

Table 2. Security feature comparisons.

Scheme	Data Confidentiality	Resistance against Man-in-the-Middle Attacks	Forward/Backward Secrecies	Filtering of Abnormal Data	Finding the Source of Abnormal Data
AMDA [10]	✓	×	×	×	×
DMDA [26]	✓	✓	×	×	×
LPSDA [21]	✓	×	×	×	×
ESPDA [27]	✓	✓	×	×	×
ERDA [28]	✓	×	✓	×	×
LCEDA [22]	✓	✓	✓	×	×
The proposed protocol	✓	✓	✓	✓	✓

7. Performance Analysis

In this section, we evaluate the performance of our scheme and compare our scheme with two representative and related schemes, the LCEDA scheme by Su et al. [20] and the DMDA scheme by Song et al. [25]. All of these schemes involve the use of masking values to encrypt the electricity usage data, and our scheme uses matrix encryption to filter abnormal data beyond that. Hence, we primarily evaluate the performance of the proposed scheme with LCEDA and DMDA in terms of communication and computation costs. Table 3 lists some notations for the performance comparisons.

Table 3. Notations.

Notation	Semantics	Notation	Semantics
$T_{(t-1)-poly}$	Time of evaluation operation of a $(t-1)$-polynomial	$\|Z_p\|$	Element size in Z_p
$\|ID\|$	Bit length of the identifier	$\|G\|$	Element size in G
$\|M_1\|$	$(2N+4) \times (2N+4)$ Matrix size	$\|M_3\|$	$1 \times (2N+4)$ Matrix size
$\|M_2\|$	$(N+5) \times (N+5)$ Matrix size	$\|M_4\|$	$(N+5) \times 1$ Matrix size
T_a	Time of an addition operation in Z_p	$\|T_h\|$	Time of a hash operation
T_s	Time of a subtraction operation in Z_p	$\|T_{pm}\|$	Time of a point multiplication operation
M_m	Time of a multiplication operation in matrix	$\|M_a\|$	Time of an addition operation in matrix

7.1. Communication Costs

The communication costs of the LCEDA, the DMDA, and our scheme in the enrollment stage are shown in Table 4. The highest communication costs are mainly concentrated between the cloud server and the smart meters in these schemes.

It costs $|Z_p| + |ID|$ communication overheads for the aggregation center to register at the cloud server in LCEDA. In addition, each smart meter spends $t|Z_p| + |ID|$ and $|Z_p|$ on registering at the cloud server and the aggregation center, respectively. Hence, in the enrollment stage, the complexity of communication times in LCEDA is $O(1)$, and the total costs are $(t+2)|Z_p| + 2|ID|$. In DMDA, the complexity of communication times is $O(1)$, and the aggregation center spends $|G|$ communication overheads on registering at the cloud server to obtain the mask values, while the smarts register spends $(t+2)|Z_p| + |G| + 2|ID|$ communication overheads. Therefore, the total length of a communication message is constant in the enrollment stage of DMDA. In our scheme, the complexity of communication times is $O(1)$, and it costs $|M_1| + |M_2| + |Z_p| + |ID|$ communication overheads for the smart meters to register at the cloud server.

To sum up, in the enrollment stage, the total length of the communication message of LCEDA in the enrollment stage is linear with t, hence, it has the highest communication costs among these schemes. Although both the DMDA and our scheme are constant, our scheme is less efficient than DMDA, comprehensively considering communication times and message length.

Table 4. Comparisons of communication costs in the enrollment stage.

Scheme	CS↔SM	AC↔SM	CS↔AC	Total Costs
LCEDA	$t\|Z_p\| + \|ID\|$	$\|Z_p\|$	$\|Z_p\| + \|ID\|$	$(t+2)\|Z_p\| + 2\|ID\|$
DMDA	$2\|Z_p\| + \|G\| + 2\|ID\|$	–	$\|G\|$	$2(\|Z_p\| + \|G\| + \|ID\|)$
The proposed protocol	$\|M_1\| + \|M_2\| + \|Z_p\| + \|ID\|$	–	$\|M_3\| + \|M_4\|$	$\|M_1\| + \|M_2\| + \|Z_p\| + \|ID\| + \|M_3\| + \|M_4\|$

7.2. Computation Costs

To evaluate performance, we conducted some experiments on a computer running Windows 10 with a 3.00 GHz Intel Core i5-8500 CPU and 8 GB memory. These experiments were run separately 50 times to obtain the mean results using the GNU Multiple Precision Arithmetic (GMP) Library and Pairing-Based Cryptography (PBC) Library.

The system initialization stage consists of two stages: the system setup stage and the enrollment stage. We set the number of users as 1000 in the implementation. The system setup stage in LCEDA, DMDA, and our scheme costed 8.74 ms, 29.9 ms, and 4.90 ms, respectively. The comparison of computation costs related to LCEDA, DMDA, and our scheme in the enrollment stage is shown in Figure 4, where we set the number of users to vary from 100 to 1000 at an increasing interval of 100. In LCEDA, the smart meters spent $(t+1)T_{(t-1)-poly}$ on registering at the cloud server and the aggregation center without negotiating with each other. As shown in Figure 4, the computation time of LCEDA ranged from 502.2 ms to 5895.2 ms when the number of users varied from 100 to 1000. The computation costs of DMDA are $2(T_{pm} + T_h + T_a)$. In our scheme, the smart meters and the aggregation center register at the cloud server, which costs $4M_m$, and the computation time of the proposed protocol ranges from 465.3 ms to 4897.6 ms.

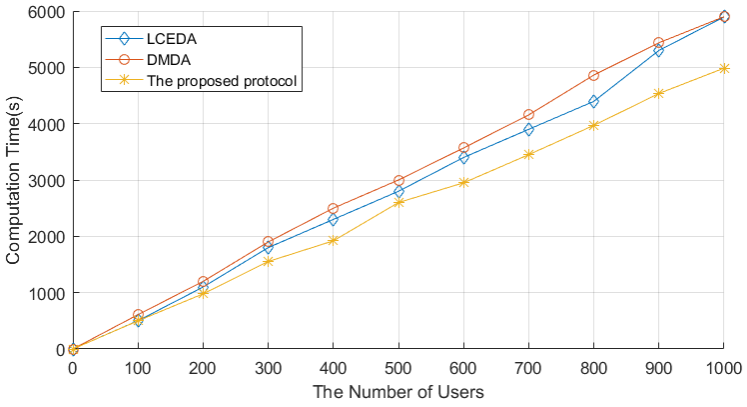

Figure 4. Computation Time of Enrollment Stage. The figure shows how the time required for LCEDA, DMDA, and the proposed scheme in the enrollment stage changes as the number of users increases. In addition, the figure reflects that the proposed scheme requires relatively little time compared to the other two.

The data collection stage consists of three stages: the data encryption stage, the aggregation stage, and the decryption stage. The encryption times of LCEDA, DMDA, and our scheme are shown in Table 5. Each smart meter in LCEDA needed 0.001 ms to encrypt the electricity usage data, while our scheme needed 0.3 ms to encrypt. The aggregation of the encrypted electricity usage data costs 28.3 ms, 28.3, and 33.2 ms in LCEDA, DMDA,

and our scheme, respectively, when the number of users is 1000. In LCEDA and DMDA, the cloud center needs 0.53 ms to decrypt the aggregation result, whereas our scheme only needs 0.13 ms. Finally, the time to encrypt data and to aggregate data in our scheme are shown in Figure 5 and Figure 6, respectively. Therefore, LECDA and DMDA have lower computation cost,s $(2(T_{pm} + T_h) + (n+1)T_a + T_s)$ and $((n+t)_{T(t-1)-ploy} + 2(n-1)T_m + (2n+1)T_a + T_s)$, respectively, compared to our scheme, which is because they do not involve filtering abnormal electricity usage data and do not support finding out the source of abnormal electricity usage data.

Table 5. Comparisons of computation costs.

Scheme (n = 1000)	Encryption		Aggregation		Decryption	
	Times (ms)	Costs	Times (ms)	Costs	Times (ms)	Costs
LCEDA	T_a	0.001	$(n-1)T_a$	28.3	T_s	0.53
DMDA	T_a	0.001	$(n-1)T_a$	28.3	T_s	0.53
The proposed protocol	$T_a + 2M_m$	0.1	$2nM_m + (n-1)T_a$	33.2	T_s	0.13

To sum up, our scheme needs to pay more computation costs for filtering abnormal users and finding out the source of abnormal users, but the increase is not significant, that is, our scheme is indeed efficient.

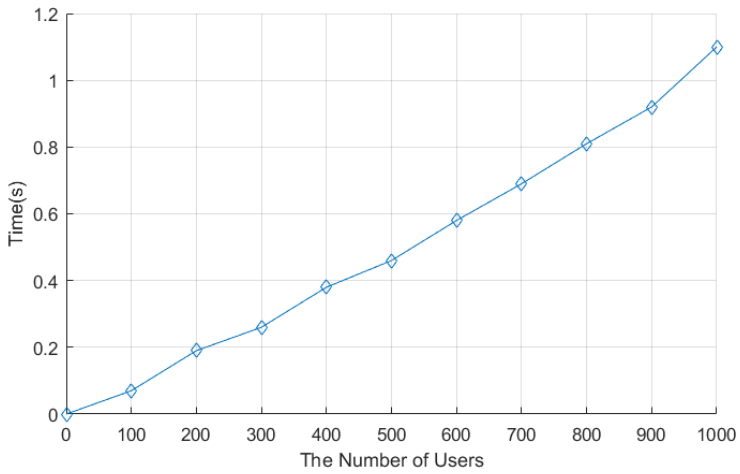

Figure 5. Time to encrypt data in the proposed protocol.

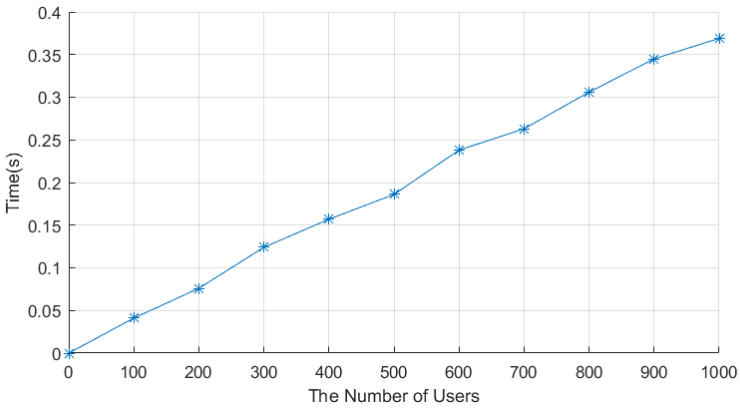

Figure 6. Time to aggregate data in the proposed protocol.

8. Conclusions

In this paper, we propose a lightweight and privacy-friendly data aggregation scheme against abnormal data to solve the problem that the abnormal electricity usage data cannot be filtered out when it is encrypted. Besides, our scheme can find out the smart meters which reported the abnormal data. Compared with other complex schemes, our scheme only uses a lightweight matrix encryption, which has lower computational costs and is more suitable for smart meters with limited computing capacity. Finally, a security analysis of our proposed scheme is presented to prove that our scheme can fully protect the privacy of users' electricity usage data. In addition, the performance evaluations and experiments validate the effectiveness and practicability of our scheme. Consequently, our scheme can be implemented in smart grids to effectively filter abnormal data and find out its source.

It is hard to say that our scheme has no drawbacks. We mainly focus on filtering abnormal data during aggregation and finding the source of the abnormal data. We use lightweight matrix encryption to process real-time electricity usage data. However, as the range of electricity usage data expands, the constructed matrix will become larger, which will gradually increase the computational and communication overheads. To overcome this problem, we mapped the original data onto a smaller data set to reduce the size of the construction matrix, and the mapping function was determined by the filtering accuracy. By sacrificing some accuracy within an acceptable range, communication overheads can be greatly reduced. In future work, we will focus on reducing computing and communication overheads while ensuring better filtering accuracy.

Author Contributions: Conceptualization, H.H. and J.Z.; methodology, H.H.; software, H.H.; validation, H.H. and J.Z.; formal analysis, H.H.; investigation, H.H.; resources, J.Z.; data curation, J.Z.; writing–original draft preparation, H.H.; writing–review and editing, H.H.; visualization, J.Z.; supervision, J.Z.; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Natural Science Foundation of Beijing (no. 4212019, M22002), the National Natural Science Foundation of China (no. 62172005), the Guangxi Key Laboratory of Cryptography and Information Security (no. GCIS201808), and the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (no. 2019BDKF JJ012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SM	smart meter
AC	aggregation center
CS	cloud server

References

1. Tian, Y.; Zhang, Z.; Xiong, J.; Chen, L.; Ma, J.; Peng, C. Achieving Graph Clustering Privacy Preservation based on Structure Entropy in Social IoT. *IEEE Internet Things J.* **2021**, *9*, 2761–2777 . [CrossRef]
2. Xiong, J.; Ma, R.; Chen, L.; Tian, Y.; Li, Q.; Liu, X.; Yao, Z. A Personalized Privacy Protection Framework for Mobile Crowdsensing in IIoT. *IEEE Trans. Ind. Inform.* **2020**, *16*, 4231–4241. [CrossRef]
3. Dominicis, S.D.; Sokoloski, R.; Jaeger, C.M.; Schultz, P.W.; Communications, P. Making the smart meter social promotes long-term energy conservation. *Palgrave Commun.* **2019**, *5*, 51. [CrossRef]
4. Onen, A.; Broadwater, R. Is the smart grid a good investment? In Proceedings of the 2015 3rd International Istanbul Smart Grid Congress and Fair (ICSG), Istanbul, Turkey, 29–30 April 2015; pp. 1–5. [CrossRef]

5. Bansal, P.; Singh, A. Smart metering in smart grid framework: A review. In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 22–24 December 2016.
6. Xiong, J.; Ren, J.; Chen, L.; Yao, Z.; Lin, M.; Wu, D.; Niu, B. Enhancing Privacy and Availability for Data Clustering in Intelligent Electrical Service of IoT. *IEEE Internet Things J.* **2019**, *6*, 1530–1540. [CrossRef]
7. Xiong, J.; Bi, R.; Zhao, M.; Guo, J.; Yang, Q. Edge-Assisted Privacy-Preserving Raw Data Sharing Framework for Connected Autonomous Vehicles. *IEEE Wirel. Commun.* **2020**, *27*, 24–30. [CrossRef]
8. Shinde, S.B.; Katti, P.K. Smart measurement techniques for efficient operation of smart grid. In Proceedings of the International Conference on Circuit, Nagercoil, India, 19–20 March 2015; pp. 1–6.
9. Vilas, V.G.; Pujara, A.; Bakre, S.M.; Muralidhara, V. Implementation of metering practices in smart grid. In Proceedings of the International Conference on Smart Technologies & Management for Computing, Chennai, India, 6–8 May 2015.
10. Hua, S.A.; Yl, A.; Zhe, X.D.; Mzab, C. An efficient aggregation scheme resisting on malicious data mining attacks for smart grid. *Inf. Sci.* **2020**, *526*, 289–300.
11. Xue, K.; Yang, Q.; Li, S.; Wei, D.; Peng, M.; Memon, I.; Hong, P. PPSO: A Privacy-Preserving Service Outsourcing Scheme for Real-Time Pricing Demand Response in Smart Grid. *IEEE Internet Things J.* **2019**, *6*, 2486–2496. [CrossRef]
12. Saleem, A.; Khan, A.; Malik, S.; Pervaiz, H.; Malik, H.; Alam, M.; Jindal, A. FESDA: Fog-Enabled Secure Data Aggregation in Smart Grid IoT Network. *IEEE Internet Things J.* **2020**, *7*, 6132–6142. [CrossRef]
13. Ara, A.; Al-Rodhaan, M.; Tian, Y.; Al-Dhelaan, A. A Secure Privacy-Preserving Data Aggregation Scheme Based on Bilinear ElGamal Cryptosystem for Remote Health Monitoring Systems. *IEEE Access* **2017**, *5*, 12601–12617. [CrossRef]
14. Liu, Y.; Guo, W.; Fan, C.I.; Chang, L.; Cheng, C. A Practical Privacy-Preserving Data Aggregation (3PDA) Scheme for Smart Grid. *IEEE Trans. Ind. Inform.* **2019**, *15*, 1767–1774. [CrossRef]
15. Romdhane, R.B.; Hammami, H.; Hamdi, M.; Kim, T.H. At the cross roads of lattice-based and homomorphic encryption to secure data aggregation in smart grid. In Proceedings of the 2019 15th International Wireless Communications Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 1067–1072. [CrossRef]
16. Abdallah, A.; Shen, X.S. A Lightweight Lattice-Based Homomorphic Privacy-Preserving Data Aggregation Scheme for Smart Grid. *IEEE Trans. Smart Grid* **2018**, *9*, 396–405. [CrossRef]
17. Kaur, K.; Garg, S.; Kaddoum, G.; Gagnon, F.; Guizani, M. A Secure, Lightweight, and Privacy-Preserving Authentication Scheme for V2G Connections in Smart Grid. In Proceedings of the International Workshop on Hot Topics in Social and Mobile Connected Smart Objects (HotSALSA'19) in Conjunction with IEEE INFOCOM 2019, Paris, France, 29 April–2 May 2019.
18. Zuo, X.; Li, L.; Peng, H.; Luo, S.; Yang, Y. Privacy-Preserving Multidimensional Data Aggregation Scheme Without Trusted Authority in Smart Grid. *IEEE Syst. J.* **2021**, *15*, 395–406. [CrossRef]
19. Ming, Y.; Zhang, X.; Shen, X. Efficient Privacy-Preserving Multi-Dimensional Data Aggregation Scheme in Smart Grid. *IEEE Access* **2019**, *7*, 32907–32921. [CrossRef]
20. Chen, Y.; Martínez-Ortega, J.F.; Castillejo, P.; López, L. A Homomorphic-Based Multiple Data Aggregation Scheme for Smart Grid. *IEEE Sens. J.* **2019**, *19*, 3921–3929. [CrossRef]
21. Gope, P.; Sikdar, B. Lightweight and Privacy-Friendly Spatial Data Aggregation for Secure Power Supply and Demand Management in Smart Grids. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1554–1566. [CrossRef]
22. Su, Y.; Li, Y.; Li, J.; Zhang, K. LCEDA: Lightweight and Communication-Efficient Data Aggregation Scheme for Smart Grid. *IEEE Internet Things J.* **2021**, *8*, 15639–15648. [CrossRef]
23. Huang, C.; Wang, X.; Gan, Q.; Huang, D.; Yao, M.; Lin, Y. A lightweight and fault-tolerable data aggregation scheme for privacy-friendly smart grids environment. *Cluster Comput.* **2021**, *24*, 3495–3514. [CrossRef]
24. Ni, J.; Zhang, K.; Alharbi, K.; Lin, X.; Zhang, N.; Shen, X.S. Differentially Private Smart Metering With Fault Tolerance and Range-Based Filtering. *IEEE Trans. Smart Grid* **2017**, *8*, 2483–2493. [CrossRef]
25. Zheng, Y.; Lu, R.; Guan, Y.; Shao, J.; Zhu, H. Towards Practical and Privacy-Preserving Multi-dimensional Range Query over Cloud. *IEEE Trans. Dependable Secur. Comput.* **2021**. [CrossRef]
26. Song, J.; Liu, Y.; Shao, J.; Tang, C. A Dynamic Membership Data Aggregation (DMDA) Protocol for Smart Grid. *IEEE Syst. J.* **2020**, *14*, 900–908. [CrossRef]
27. Sui, Z.; Meer, H.d. An Efficient Signcryption Protocol for Hop-by-Hop Data Aggregations in Smart Grids. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 132–140. [CrossRef]
28. Xue, K.; Zhu, B.; Yang, Q.; Wei, D.S.L.; Guizani, M. An Efficient and Robust Data Aggregation Scheme Without a Trusted Authority for Smart Grid. *IEEE Internet Things J.* **2020**, *7*, 1949–1959. [CrossRef]