Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.elsevier.com/locate/dcan

Achieving privacy-preserving big data aggregation with fault tolerance in smart grid

Zhitao Guan [*], Guanlin Si

School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China

ARTICLE INFO

Keywords:
Big data
Smart grid
Privacy-preserving
Fault tolerance

ABSTRACT

In a smart grid, a huge amount of data are collected for various applications, such as load monitoring and demand response. These data are used for analyzing the power state and formulating the optimal dispatching strategy. However, these big energy data in terms of volume, velocity and variety raise concern over consumers' privacy. For instance, in order to optimize energy utilization and support demand response, numerous smart meters are installed at a consumer's home to collect energy consumption data at a fine granularity, but these fine-grained data may contain information on the appliances and thus the consumer's behaviors at home. In this paper, we propose a privacy-preserving data aggregation scheme based on secret sharing with fault tolerance in a smart grid, which ensures that the control center obtains the integrated data without compromising privacy. Meanwhile, we also consider fault tolerance and resistance to differential attack during the data aggregation. Finally, we perform a security analysis and performance evaluation of our scheme in comparison with the other similar schemes. The analysis shows that our scheme can meet the security requirement, and it also shows better performance than other popular methods.

1. Introduction

As a new generation of energy networks, the smart grid is considered a useful means of solving the severe environmental issues and resource crisis. It is the product of the combination of energy network and information technology. Compared to the unidirectional centralized grid, the control mode of the smart grid is more flexible and reliable. It supports bidirectional power flow between the user and the grid. A user in a smart grid is not only a consumer but also a generator. In a smart grid, large quantities of data are collected to support basic services [1]. For example, to create a power plan or dynamic pricing, the control center needs to collect and analyze real-time data from various applications by adopting a smart meter (SM) installed at the user's house. Furthermore, electric vehicle drivers need to upload their location message to the control center for power dispatching.

Although big data collected from users is necessary for the basic service, it is usually sensitive to users' privacy [2]. For instance, SMs are adopted for the control center to collect the real-time data from users, but these data may disclose the users' behaviors. Thus, if a thief obtains the real-time data from user's SM, they may gain entry to the user's house when they notice that nobody is at home. Besides, a user's location-privacy may be disclosed during the interaction between electric vehicles and a smart grid, which may help an adversary catch the user [3,4]. If a user's sensitive data is not well protected, the implementation of the smart grid will meet resistance. Therefore, privacy leakage in smart grid is an extremely important problem.

There are various solutions for privacy-preservation in a smart grid. As we know, conventional privacy-preserving strategies can be divided into two. One is to hide the user's identity, while the other is to protect the user's sensitive data [5]. As the big data in a smart grid are characterized by volume, velocity and variety [6], approaches for privacy-preservation need to focus more on communication overhead and computational cost [7,8].

The first strategy is to protect users' identities through anonymity or pseudonyms. Users' attributes can be classified into identity information, quasi-identifiers, and sensitive information. Given an anonymity table, if the attributes in the table have not been properly treated, an adversary may deduce the relationship between a user's identity and sensitive information according to the user's quasi-identifiers such as age and gender. Although the k-anonymity algorithm [9] and l-diversity

[*] Corresponding author.
E-mail addresses: guanzhitao@126.com, m18811612766@163.com (Z. Guan).

https://doi.org/10.1016/j.dcan.2017.08.005
Received 15 February 2017; Received in revised form 3 August 2017; Accepted 10 August 2017
Available online xxxx

algorithm [10] address the disadvantages of the identity-protection scheme, it is very difficult to find a credible party to complete the secure anonymity work.

The second strategy is to use data aggregation to protect users' real-time data, which includes homomorphic encryption and data-obfuscation. In fact, these two methods are often used together. Given the huge volume, velocity and variety of big data, we aggregate users data in groups and adopt the Paillier algorithm to encrypt users' real-time data. In addition, secret shares are distributed to each user for further obfuscating their data. Our contributions are summarized as follows: 1) In case the data aggregation device (DA) and control center (CC)launch a differential attack based on two data sets differing in at most one element, the threshold of secret shares is the same as the number of group members. If the number of SMs participating in data aggregation is not equal to the number of group members, CC cannot obtain the correct result.

2) We mask users' identities through anonymity and use the group's hash table to search for the malfunctioning SM through comparison with other groups without disclosing users' identities.
3) We realize the fault tolerance through the substitution strategy. Each member in a group has the same secret share as other group members. When there is a malfunctioning SM in a group, we can use related users' data in other groups to substitute.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the system model and design goals. In Section 4, some preliminaries are presented. In Section 5, our scheme is stated. In Section 6, security analysis is presented. In Section 7, the performance of our scheme is evaluated. In Section 8, the paper is concluded.

2. Related work

To protect the privacy of users in a smart grid, many scholars have proposed various strategies. These strategies can be classified into two: 1) protect users' privacy by masking the real identities, and 2) protect users' privacy by masking their real-time data.

Some works are focused on masking users' identity. A simple solution adopting a trusted-party to manage the identity list was proposed in Ref. [11]. However, finding a trusted-party is not easy. Privacy preservation and validity-authentication are two related problems [12] and Cheung proposed a scheme based on blind signatures to solve these two problems in Ref. [13]. It ensures that the verifier can authenticate the sender's signature without access to their private information. Camenisch and Lysyanskaya proposed a scheme named CL -Signature scheme which is similar to the blind signature proposed in Ref. [14]. Stegelmann proposed k-anonymity to protect users' privacy [15], but finding a credible party to complete the secure anonymity work is difficult. An effective scheme based on a virtual ring was presented in Ref. [16]. It groups users by their geographical positions and distributes the same serial number to each member in the same group. In this manner, the control center can obtain all of the users' data without knowing the senders' real identities. However, validity-authentication cannot be guaranteed because of the anonymity. Riesch, analyzed the authentication problems occuring during identity preservation [17]. A privacy-preserving scheme based on pseudonyms is also very common, as reported in [18–20], and it is always combined with the ring signature or blind signature to mask users' identity.

Some works focused on masking users' real-time data. Solutions using a battery to hide the real-time data were proposed in Ref. [21,22]. In this type of scheme, a smart grid and household battery simultaneously provide users with electricity. When the household consumption becomes high, the battery discharges. Otherwise, it charges. In this manner, the user's real-time data can be hidden to protect privacy. The disadvantage of this method is that the frequent charging and discharging of the battery may reduce the lifetime of the battery. Data aggregation is also popular in smart grids for privacy preservation. Paillier encryption [23–25] and Bone-Goh-Nission encryption [26,27] are classical homomorphic encryption algorithms for data aggregation. In addition, bilinear mapping is a common solution for data aggregation [28,29]. Remarkably, it is often used to realize key-exchange. A secret sharing scheme was proposed to realize the data aggregation. The Shamir technique is often adopted to encrypt the electricity data [30,31]. It divides a secret into different pieces and distributes them to various entities in a smart grid. The control center can obtain the integral secret only if it acquires fixed number of shares. Of course, the secret sharing scheme can be also applied to other aspects, such as key management, but this property is not taken into our consideration, and we will describe it in detail below. Beussink, reported a scheme based on data obfuscation [32], which adds a random number to each electricity data, but it will cause some large errors if the random numbers are not reasonable.

Differential privacy [33] and fault-tolerance are two important problems occuring during the data aggregation and data obfuscation. In Ref. [23], Shi and Sun discussed fault tolerance and differential privacy. However, the error rate and computational complexity are not very ideal. For the differential privacy, scholars always add random noise that obeys Laplace distribution. However, adding random noise into users' data will cause unnecessary errors. Hong also analyzed the tradeoff between differential privacy and utility [34].

In our scheme, we used substitution to realize fault tolerance based on secret sharing during the data aggregation. Compared with other fault tolerance schemes such as [23], our scheme has a lower error rate and less computational cost. In addition, we use a secret sharing scheme to defend against differential attack. The process of our scheme is shown in Fig. 1.

3. System model and design goals

3.1. System model

As shown in Fig. 1, a smart grid is divided into four parts: the control center (CC), the key initialization center (KIC), the data aggregation device (DA), and residential users.

1) Residential users: We divide all the users into different groups in accordance with their geographical locations. Each residential user's house is equipped with a smart meter (SM) to collect the real-time data of the house appliances every 15 min.
2) Data aggregation device: The data aggregation device is responsible for collecting all the data sent by SMs, calculating the sum of ciphertexts by running the homomorphic algorithm and uploading the sum to the control center. In addition, it has a fault-tolerance function. When an SM is malfunctioning in a group, its data would be replaced by another group's SM which has the same secret key.
3) Key initialization center: The key initialization center is responsible for initializing all of the keys for SMs and CC. Additionally, each group has the same number of SMs, and the encrypted parameters assigned to each group are equal.
4) Control center: The control center can acquire a summary of real-time data in a smart grid from DA by a decrypted key. With these data, the CC can obtain the trend of power consumption and create a power generation plan or dynamic pricing immediately.

3.2. Adversarial model

We assume that the smart meter installed on the user side is vulnerable to external attacks. The communication channel is not secure and an

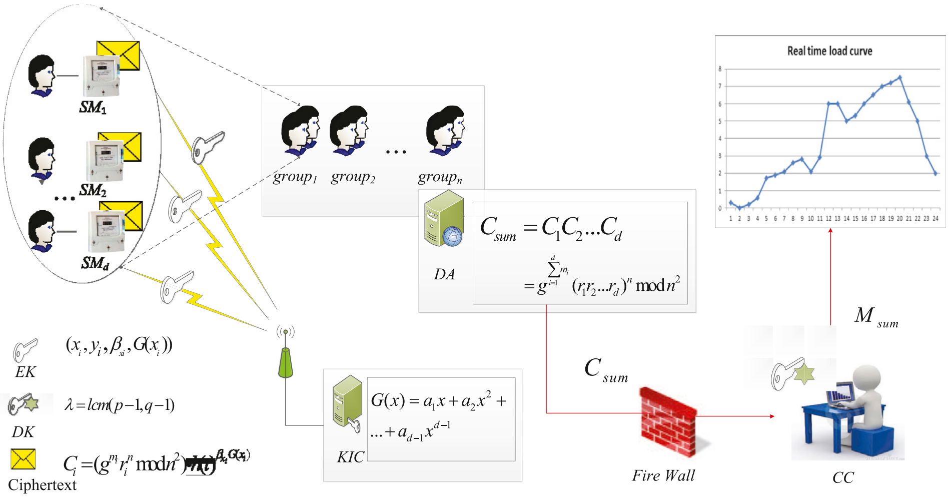
SM1
SM2
SMd
group1
group2
groupn
DA
$C_{sum} = C_1C_2...C_d$
$= g^{\sum_{i=1}^{d} m_i} (r_1r_2...r_d)^n \bmod n^2$
Real time load curve
EK
$(x_i, y_i, \beta_{xi}, G(x_i))$
DK
$\lambda = lcm(p-1, q-1)$
Ciphertext
$C_i = (g^{m_i} r_i^n \bmod n^2) h(t)^{\beta_{x_i} G(x_i)}$
KIC
$G(x) = a_1x + a_2x^2 + ... + a_{d-1}x^{d-1}$
C_{sum}
Fire Wall
M_{sum}
CC

Fig. 1. System model.

adversary may eavesdrop on the channel. CC and DA are honest-but-curious. That is, they do not destroy or modify users' data, but they always attempt to acquire users' private information through the background knowledge. Moreover, CC may conspire with DA to increase the probability of a successful attack.

3.3. Design goals

Considering the above scenarios, our design goals can be divided into three aspects.

1) Privacy preservation:

A residential user's data are inaccessible to other users. Irrespective of the external adversary, DA or CC should not acquire the real-time data of users even if it knows the cipher text and encryption algorithms.

2) Resistance to differential attack:

Although the plaintext of a single user can be masked by homomorphic encryption after the data aggregation, given two data sets differing in at most one element, an adversary can obtain a user's plaintext by calculating the difference between the two data sets. We call this attack method differential attack. Therefore, resistance to such differential attacks is one of our design goals.

3) Fault tolerance:

As we set the threshold of secret shares to be the same as the number of group members, only if DA aggregates all the ciphertexts in a group and sends them to CC, can CC obtain the correct result. Therefore, if there is a damaged SM, the data aggregation cannot run correctly. Fault tolerance stipulates that the data aggregation must continue to run normally, when there are several malfunctioning SMs in a group.

4. Preliminaries

4.1. Notations

In Table 1, the notations used in the proposed scheme are listed.

Table 1
Notations.

Acronym	Descriptions
SM	Smart meter
CC	Control center
EK	Encrypted key
DK	Decrypted key
KIC	Key initialization center
DA	Data aggregation device
SNR	Signal-to-noise ratio
gcd	Greatest common denominator
lcm	Least common multiple
P_S	The power producing the normal signal
P_N	The power producing the noise
m_i	Plaintext of SM_i
r	Random number
C_i	Ciphertext of SM_i
d	The number of group members
x_i	The serial number of SM_i
y_i	The serial number of $group_i$
t	Time stamp
e%	The error rate
$\tilde{m}$	The average value of plaintext
$\tilde{C}$	The average value of ciphertext
N	The total number of all the SMs
F	The number of the malfunctioning SMs
M_{sum}	The sum of users' plaintexts
C_{sum}	The sum of users' ciphertexts
C_{sum3}	The sum of lost ciphertexts
$C\prime_{sum}$	The sum of processed ciphertexts

4.2. Paillier cryptosystem

The Paillier encryption algorithm is an asymmetric encryption algorithm, which has additive homomorphism properties. It includes three procedures: key generation, encryption and decryption.

1) Key generation: We choose two prime numbers p, q with the same length and calculate $n = pq$. g is a generator of cyclic group $Z^*_{n^2}$, and $gcd(L(g^\lambda modn^2), n) = 1$. The public key is (n, g), and the private key is λ.

$$\lambda = lcm(p-1, q-1) \tag{1}$$

2) Encryption phase: For the plain-text $m \in Z_n$, we can select a random number $r < n$. Then, the ciphertext can be calculated as follows:

$$C = g^m r^n mod n^2 \tag{2}$$

3) Decryption phase: After receiving the cipher-text C, the receiver can obtain the plain text m with the private key λ by using the following equation:

$$m = \frac{L(C^{\lambda} mod n^2)}{L(g^{\lambda} mod n^2)} mod n \tag{3}$$

4.3. Secret sharing scheme

The secret sharing scheme is a scheme that splits a secret into α pieces and distributes these pieces with different valid members. If an adversary captures a member in the system, they can only obtain a piece of the secret. Only if the adversary obtains at least d pieces of the secret, can they obtain the whole secret. We call d the threshold and usually adopt the Shamir technique to realize this result.

The trusted-party chooses a polynomial to split a secret.

$$G(x) = \theta + a_1 x + a_2 x^2 + \ldots + a_d x^d \tag{4}$$

where $(x_i, G(x_i))$ is the corresponding share. Remarkably, the Shamir secret sharing scheme is the fully homomorphic and can be designed as a better scheme to realize the data aggregation. According to the Lagrange interpolation polynomial, we have

$$G(x) = \sum_{j=1}^{d} \left(\prod_{i=1, i \neq j}^{d} \frac{x_i - x}{x_i - x_j} \right) \tag{5}$$

$$\beta_{x_i} = \prod_{j \neq i}^{d} \frac{x_j}{x_j - x_i} \tag{6}$$

Then, we can easily compute θ as follows:

$$\sum_{i=1}^{d} G(x_i) \beta_{x_i} = \theta = 0 \tag{7}$$

4.4. Signal-to-noise ratio

The signal-to-noise ratio (SNR) is a common quantity that is often used to measure the performance of electronic systems. SNR is calculated as follows:

$$SNR = 10 lg \frac{P_S}{P_N} \tag{8}$$

where P_S represents the power producing the normal signal and P_N represents the power producing the noise. The higher is the SNR, the stronger is the signal. In general, when the image SNR is greater than 30 dB, the resolution of the picture is not affected. In this paper, we take the sum of normal SMs'data as the signal and take the sum of malfunctioning SMs'data processed by substitution as the noise. Then, we can measure the error rate of our scheme through the image SNR.

5. Our scheme

5.1. System initialization

The KIC first chooses two prime numbers p and q with the same length and calculates $n = pq$. g is the generator of the cyclic group $Z_{n^2}^*$, and $gcd(L(g^{\lambda} mod n^2), n) = 1$. The public key is (n, g), and the private key is $\lambda = lcm(p-1, q-1)$.

It constructs a formula $G(x) = \theta + a_1 x + a_2 x^2 + \ldots + a_d x^d$. Here, we set $\theta = 0$. All the SMs in the smart grid are divided into different groups, and each group has d members. For each SM in a group, the KIC assigns the private key $[x_i, y_i, G(x_i), \beta_{x_i}]$ to the SM_i, where x_i is a random number representing the SM_i serial number, y_i is the group serial number, and $\beta_{x_i} = \prod_{j \neq i}^{d} \frac{x_j}{x_j - x_i}$. Particularly, the set of SM serial numbers $[x_1, x_2, \ldots, x_d]$ in each group is the same. That is, a special SM can find members by using the same private key except for the group serial number in other groups.

Finally, KIC publishes (n, h, g, H_1, H_2) and sends λ to CC securely. h, H_1, and H_2 are three hash functions. **Algorithm 1** shows the process of system initialization (see Table 2).

5.2. Encryption

The SM collects the electricity data every 15 min from all the house applications. For time t, it computes

$$C_i = \left(g^{m_i} r_i^n mod n^2\right) h(t)^{\beta_{x_i} G(x_i)} \tag{9}$$

$$H_1\left(t \middle| G(x_i) \beta_{x_i}\right) \tag{10}$$

$$H_2\left(y_i \middle| C_i \middle| H_1\left(t \middle| G(x_i) \beta_{x_i}\right)\right) \tag{11}$$

Then, the SM encrypts the total electricity data and sends y_i, C_i, $H_1(t|G(x_i)\beta_{x_i})$, and $H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$ to the DA. Remarkably, a user's identity can be confirmed through (x_i, y_i). The publishing of the group serial number y_i while masking the serial number x_i can still protect user's identity. **Algorithm 2** shows the process of encryption (see Table 3).

5.3. Data aggregation

When the DA receives a message from an SM, it verifies $H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$ to authenticate the message integrity. If the hash value is correct, data aggregation will be performed.

1) Normal aggregation: If the DA receives all the SMs' data in a group, it runs the data aggregation as follows:

$$\begin{aligned} C_{sum} &= \prod_{i=1}^{d} C_i \\ &= \left(g^{\sum_{i=1}^{d} m_i} \left(\prod_{i=1}^{d} r_i \right)^n mod n^2 \right) h(t)^{\sum_{i=1}^{d} \beta_{x_i} G(x_i)} \\ &= g^{\sum_{i=1}^{d} m_i} \left(\prod_{i=1}^{d} r_i \right)^n mod n^2 \end{aligned} \tag{12}$$

Table 2
System initialization.

```
Algorithm1: System initialization

KIC.Input: $d$
KIC.Output:$[x_i, y_i, G(x_i), \beta_{x_i}]$,$(n, h, g, H_1, H_2)$, $\lambda$

(1) Choose two primes $p$ and $q$ and calculate $n = pq$
(2) Calculate $\lambda = lcm(p-1, q-1)$ and send to CC
(3) Choose $g \in \{Z_{n^2}^* | gcd(L(g^{\lambda} mod n^2), n) = 1\}$
(4) Construct $G(x) = \theta + a_1 x + a_2 x^2 + \ldots + a_d x^d$
(5) Set $x_i$ for each SM and set $y_i$ for each group
(6) Calculate $G(x_i)$ and $\beta_{x_i}$
(7) Send $[x_i, y_i, G(x_i), \beta_{x_i}]$ to each SM
(8) Choose hash functions $h, H_1, H_2$
(9) Publish $(n, h, g, H_1, H_2)$
```

Table 3
Encryption.

```
Algorithm2: Encryption
SM.Input: $[x_i, y_i, G(x_i), \beta_{x_i}]$,$(n, h, g, H_1, H_2)$
SM.Output:$y_i$,$C_i$,$H_1(t|G(x_i)\beta_{x_i})$,$H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$
(1) Choose $r_i = GenRan() \cap \{r_i < n\}$
(2) Calculate $C_i = (g^{m_i} r_i^n modn^2) h(t)^{\beta_{x_i} G(x_i)}$
(3) Calculate $H_1(t|G(x_i)\beta_{x_i})$ and $H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$
(4) send $y_i$, $C_i$, $H_1(t|G(x_i)\beta_{x_i})$,
$H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$ to DA
```

The DA calculates the hash value of the concatenation of C_{sum} and t, and encrypts the value by using his private key sk_D. We show the final result as follows:

$$Enc[sk_D, H_1(C_{sum}|t)] \tag{13}$$

2) Fault tolerance: As DA only recognizes the group that an SM belongs to according to the group serial number, we can use $H_1(t|G(x_i)\beta_{x_i})$ to find the malfunctioning SM while masking user's identity. If a malfunctioning SM exists in a group, the DA runs the following steps:

First, the DA compares the hash table of this group constituted by $H_1(t|G(x_i)\beta_{x_i})$ with other complete groups to find the malfunctioning SM.

Then, the DA selects SM_j from other groups with the same hash value $H_1(t|G(x_i)\beta_{x_i})$ to replace the malfunctioning SM_i. Theoretically, if a malfunctioning SM exists, we should not consider this user's data. To further reduce the error, the data of SM_j is processed before the data aggregation as follows:

$$\tilde{m} = \frac{\sum_{i=1}^{d} m_i}{d} \tag{14}$$

$$C_j' = \frac{C_j}{g^{\tilde{m}}} = \left(g^{m_j - \tilde{m}} r_j^n modn^2\right) h(t)^{\beta_{x_j} G(x_j)} \tag{15}$$

where $\sum_{i=1}^{d} m_i$ represents the sum of the electricity data of the previous period in $group_j$. C_j' represents the processed data of C_j and replaces the missing data C_i to run data aggregation. **Algorithm 3** shows the process of data aggregation (see Table 4).

Table 4
Data aggregation.

```
Algorithm3: Data aggregation
DA.Input: $y_i$,$C_i$,$H_1(t|G(x_i)\beta_{x_i})$,$H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$
DA.Output:$C_{sum}$, $Enc[sk_D, H_1(C_{sum}|t)]$
(1) Calculate $H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$ based on input.
(2) If $H_2(y_i|C_i|H_1(t|G(x_i)\beta_{x_i}))$ is right, then
(3) For $1 \le i \le v$
(4) Calculate the number of SM in $group_i$
(5) If(Count(SM) $= d$), then
(6) $C_{sum} = \prod_{i=1}^{d} C_i$
(7) $i++$, return to (3)
(8) else
(9) Compare hash table composed of
$H_1(t|G(x_i)\beta_{x_i})$ with other groups
(10) If there is a value lost in $group_i$ , then
(11) Calculate $\tilde{m}$ and $C_j'$, return to (6)
(12) else if there is an extra value in $group_i$
(13) Drop this message
(14) end if
(15) end if
(16) end if
```

5.4. Power dispatching

After receiving the aggregated result from the DA in the smart grid, CC first uses the public key of the DA to decrypt the value of $Enc[sk_D, H_1(C_{sum}|t)]$, which is used to ensure that the message is from the DA. Then, it calculates the hash value of the concatenation of C_{sum} and t. If the final result is the same as the attached one, CC can calculate the sum of the electricity data by using the decrypted key λ as follows:

$$M_{sum} = \frac{L\left(C_{sum}^{\lambda} modn^2\right)}{L\left(g^{\lambda} modn^2\right)} modn \tag{16}$$

where M_{sum} denotes the sum of users' plaintexts. Based on the sum of the real-time data, CC can draw the real-time load curve and create the dynamic pricing, power generation plan, and other scheduling strategies. **Algorithm 4** shows the process of power dispatching (see Table 5).

6. Security analysis

6.1. Privacy-preserving

For an SM in a smart grid, we can analyze the security of data according to the following aspects: external attacker, DA, CC, and conspiracy attack.

1) External attacker: When an external attacker compromises the user's SM, cipher text C_i sent by the user can be obtained. However, because the attacker does not know the other $d-1$ users' private key and decrypted key λ, the attacker cannot acquire the plaintext.
2) DA: After receiving all the data from SMs, the DA can only perform data aggregation and replace the malfunctioning SM's data when necessary. However, it cannot obtain a single user's plaintext because of the lack of $h(t)^{\beta_{x_i} G(x_i)}$; therefore, the security of the user's data can be guaranteed.
3) CC: CC can only obtain all the users' aggregated data; thus, it cannot snoop into a single user's real-time data. Thus, user's privacy can be guaranteed in our scheme.
4) Conspiracy attack: If DA receives the decrypted key from CC and tries to acquire the plaintext of a single user, the user's privacy can still be preserved because the device does not know the secret shares $h(t)^{\beta_{x_i} G(x_i)}$. Moreover, if the DA is in collusion with some SMs from other groups with the same SM key, our scheme can protect the users' privacy because of anonymity.

6.2. Resistance to differential attack

Theorem 1. *If the threshold of the secret sharing scheme is the same as that of the number of group members during data aggregation, then this scheme can resist differential attack.*

Table 5
Power dispatching.

```
Algorithm4: Power dispatching
CC.Input: $C_{sum}$, $Enc[sk_D, H_1(C_{sum}|t)]$, $\lambda$
CC.Output:$M_{sum}$
(1) Decrypt $Enc[sk_D, H_1(C_{sum}|t)]$ by $pk_D$
(2) Calculate $H(C_{sum}|t)$ based on input
(3) If $H(C_{sum}|t)$ is wrong, then
(4) Drop this message
(5) else
(6) Calculate $M_{sum}$ by formula(16)
(7) end if
(8) For $1 \le i \le v$
(9) Calculate $\tilde{m}$ for $group_i$ and send to DA
(10) $i++$, return to (8)
```

Proof. 1) *Given two data sets differing on at most one element, C_{sum1} and C_{sum2}, we can describe them as follows:*

$$C_{sum1} = C_1C_2\ldots C_d \tag{17}$$

$$C_{sum2} = C_1C_2\ldots C_dC_{d+1} \tag{18}$$

2) *If we do not consider the obfuscation of secret share, and an adversary obtains the decrypted key λ from CC and the two data sets from DA, he can launch the differential attack as follows:*

$$C_i = g^{m_i}r_i^n modn^2 \tag{19}$$

$$M_{sum1} = \frac{L\left(C_{sum1}^{\lambda} modn^2\right)}{L\left(g^{\lambda} modn^2\right)} modn \tag{20}$$

$$M_{sum2} = \frac{L\left(C_{sum2}^{\lambda} modn^2\right)}{L\left(g^{\lambda} modn^2\right)} modn \tag{21}$$

$$m_d = M_{sum2} - M_{sum1} \tag{22}$$

3) *When we add secret shares to further obfuscate the data and set the threshold of the secret sharing scheme the same as the number of group members during data aggregation, we have*

$$C_i = \left(g^{m_i}r_i^n modn^2\right)h(t)^{\beta_{x_i}G(x_i)} \tag{23}$$

$$M_{sum1} = \frac{L\left(C_{sum1}^{\lambda} modn^2\right)}{L\left(g^{\lambda} modn^2\right)} modn \tag{24}$$

However, because the multiplication of $d+1$ shares $h(t)^{\beta_{x_i}G(x_i)}$ is not one, the adversary cannot obtain the value of M_{sum2}.

$$M_{sum2} \neq \frac{L\left(C_{sum2}^{\lambda} modn^2\right)}{L\left(g^{\lambda} modn^2\right)} modn \tag{25}$$

The user's plaintext m_d is secure because the adversary does not know the value of M_{sum2}.

Thus, we complete our Proof.

6.3. Fault tolerance

Theorem 2. *When we substitute the data of a malfunctioning SM with the data of a normal SM with the same secret share in other groups, the error caused by substitution can be ignored.*

Proof. 1) *We suppose that there are F malfunctioning SMs in a smart grid composed of N SMs. Then, the error rate can be calculated as follows:*

$$e\% = \frac{F\tilde{C}/g^{\tilde{m}}}{(N-F)\tilde{C}} \tag{26}$$

2) *As the number of SMs is much larger than the number of malfunctioning SMs, the error rate can be simplified as follows:*

$$e\% = \frac{1}{(N/F-1)g^{\tilde{m}}} \approx \frac{F}{Ng^{\tilde{m}}} \tag{27}$$

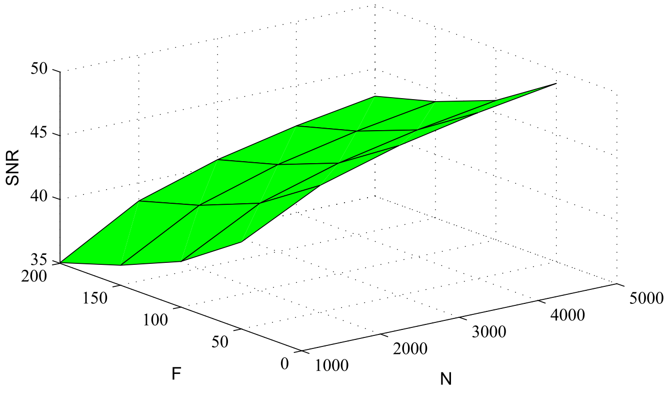

Fig. 2. Signal to noise ratio in our scheme.

3) *The larger the number of SMs, the smaller is the error rate. For large quantities of SMs in a smart grid, the error rate caused by substitution can be ignored.*

Thus, we complete our Proof.

7. Performance evaluation

The most important concept of our study is to use substitution to realize fault tolerance, thus necessitating the proof of the substitution between two members in different groups for the power dispatching in a smart grid.

Here, we take the sum of the normal SMs' data as the signal and take the sum of malfunctioning SMs' data processed by substitution as the noise. Then, we can use the SNR to measure the error rate of our scheme and the detailed formula is presented as follows.

$$SNR_{OS} = 10lg\frac{C_{sum} - C_{sum3}}{C'_{sum}} = 10lg\left(\frac{N}{F} - 1\right)g^{\tilde{m}} \tag{28}$$

By inputting the number of malfunctioning SMs and the total number of SMs in the smart grid into formula (28), where C_{sum} denotes the sum of normal SMs' data, C_{sum3} denotes the sum of malfunctioning SMs' data, and $C\prime_{sum}$ denotes the sum of the data processed by the fault tolerance, we can obtain the SNR curve, as shown in Fig. 2.

If the number of SMs is from 1000 to 5000, the SNR of our scheme is more than 35 dB in the worst case and the average rate is close to 40 dB, which is allowed in a smart grid for power scheduling. As shown in formula (28), the SNR is related to the value of g. Therefore, we can increase the accuracy by adjusting the value of g.

We use T_{exp} to denote the time of exponentiation and T_{mul} to denote the time of multiplication. T_{ran} denotes the time of generation of a

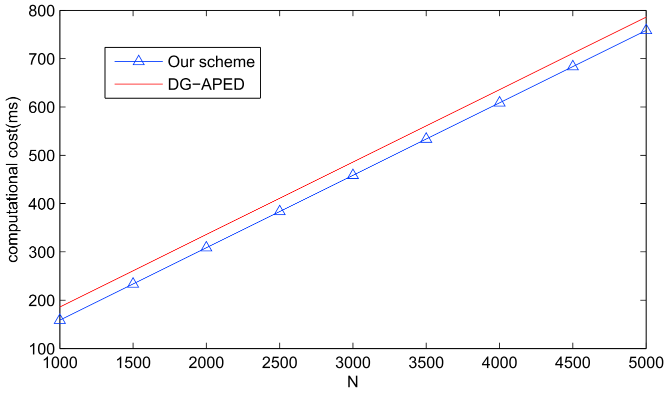

Fig. 3. Computational complexity in a normal situation.

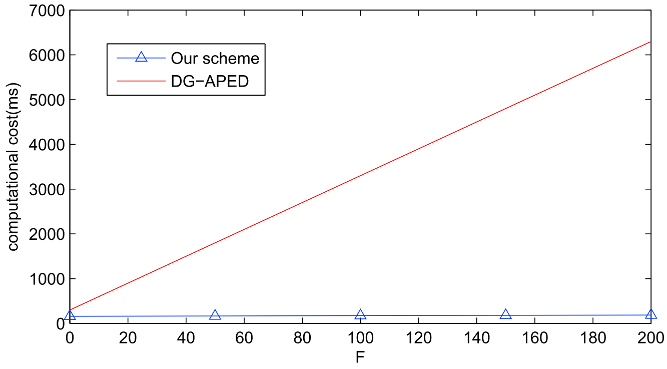

Fig. 4. Computational complexity considering fault tolerance.

random number and T_{pro} denotes the time of Pollard's lambda method. As far as we know, the DG-APED scheme proposed by Shi et al. is also a subtle group-based scheme with fault tolerance [23]. We can calculate computational cost of our scheme and DG-APED when there is no malfunctioning SM as follows:

$$T_{OS}=(N+4)T_{mul}+5T_{exp} \tag{29}$$

$$T_{DG}=(N+5)T_{mul}+7T_{exp}+T_{ran}+T_{pro} \tag{30}$$

As shown in Fig. 3 our scheme is greatly advantageous over DG-APED when the number of SM varies from 1000 to 5000.

Next, we calculated the computational cost of our scheme and DG-APED when there are malfunctioning SMs. To observe the relationship between computational cost and malfunctioning SMs, we assumed the number of SMs in the smart grid to be 1000. L denotes the number of SM types and ω denotes the number of groups in each type. k is the number of SMs in a group in the DG-APED scheme.

$$T_{OS}=(F+N+4)T_{mul}+5T_{exp} \tag{31}$$

$$T_{DG}=(N+5)T_{mul}+7T_{exp}+T_{ran}+\left(\frac{(F+1)\omega}{2}+L\right)T_{pro} \tag{32}$$

Fig. 4 shows that our scheme has a greater advantage than DG-APED when the number of malfunctioning SM varies from 0 to 200.

Furthermore, we calculated the SNRs of our scheme and DG-APED. The SNRs of the two schemes are as follows:

$$SNR_{OS}=10lg\frac{C_{sum}-C_{sum3}}{C'_{sum}}=10lg\left(\frac{N}{F}-1\right)g^{\tilde{m}} \tag{33}$$

$$SNR_{DG}=10lg\frac{2\left(N\tilde{C}-F\tilde{C}\right)}{(k-1)F\tilde{C}}=10lg\frac{2(N-F)}{(k-1)F} \tag{34}$$

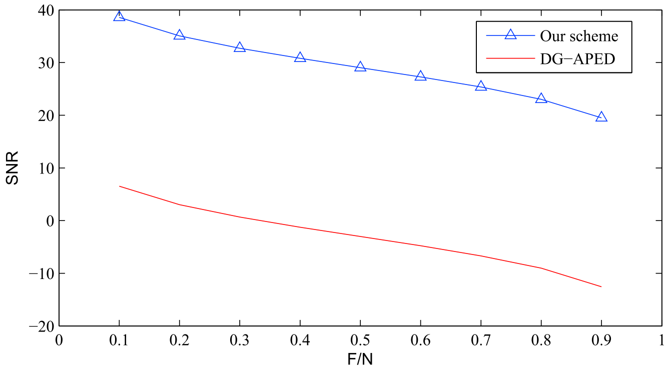

Fig. 5. Signal to noise ratio of two schemes.

Through Fig. 5, we can find that the SNR of our scheme is much higher than that of DG-APED, indicating that our scheme has less error rate.

8. Conclusion

In this paper, we proposed a privacy-preserving data aggregation scheme based on the secret sharing scheme. We set the threshold of secret shares that are the same as the number of group members to resist the differential attack by the conspiracy between DA and CC. In addition, we masked the user's identity by using the same group serial number, adopted a hash table to find the malfunctioning SM, and achieved fault tolerance for the normal aggregation by substitution. Therefore, even if there are some malfunctioning SMs in a group, our scheme can run normally. In the future, we will focus on combining the real-time data privacy with the normal billing.

Acknowledgements

This work is partially supported by Natural Science Foundation of China under grant 61402171, the Fundamental Research Funds for the Central Universities under grant 2016MS29.

References

[1] H. Zhang, Q. Zhang, Z. Zhou, X. Du, Processing geo-dispersed big data in an advanced mapreduce framework, Netw. IEEE 29 (5) (2015) 24–30.
[2] S. Yu, Big privacy: challenges and opportunities of privacy study in the age of big data, IEEE Access 4 (2016), 1–1.
[3] P. Kamat, Y. Zhang, W. Trappe, C. Ozturk, Enhancing source-location privacy in sensor network routing, in: IEEE International Conference on Distributed Computing Systems, 2005. ICDCS 2005. Proceedings, 2005, pp. 599–608.
[4] K.P.N. Puttaswamy, S. Wang, T. Steinbauer, D. Agrawal, A.E. Abbadi, C. Kruegel, B.Y. Zhao, Preserving location privacy in geosocial applications, IEEE Trans. Mob. Comput. 13 (1) (2014) 159–173.
[5] H. Zhang, Z. Xu, Z. Zhou, J. Shi, Clpp: context-aware location privacy protection for location-based social network, in: IEEE International Conference on Communications, 2015, pp. 1164–1169.
[6] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, S. Guo, Protection of big data privacy, IEEE Access 4 (2016) 1821–1834.
[7] G. Sand, L. Tsitouras, G. Dimitrakopoulos, V. Chatzigiannakis, A big data aggregation, analysis and exploitation integrated platform for increasing social management intelligence, in: IEEE International Conference on Big Data, 2015, pp. 40–47.
[8] P. Costa, A. Donnelly, A. Rowstron, et al., Camdoop: exploiting in-network aggregation for big data applications, in: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, USENIX Association, 2012, 3–3.
[9] L. Sweeney, K-anonymity, Int. J. Uncertain., Fuzziness and Knowledge-Based Syst. 10 (05) (2008) 557–570.
[10] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, in: International Conference on Data Engineering, 2006, 24–24.
[11] C. Efthymiou, G. Kalogridis, Smart grid privacy via anonymization of smart metering data, in: First IEEE International Conference on Smart Grid Communications, 2010, pp. 238–243.
[12] R. Doss, W. Zhou, S. Sundaresan, S. Yu, L. Gao, A minimum disclosure approach to authentication and privacy in rfid systems, Comput. Netw. Int. J. Comput. Telecommun. Netw. 56 (15) (2012) 3401C3416.
[13] J.C.L. Cheung, T.W. Chim, S.M. Yiu, V.O.K. Li, Credential-based privacy-preserving power request scheme for smart grid network, in: Global Telecommunications Conference, 2011, pp. 1–5.
[14] J. Camenisch, A. Lysyanskaya, Signature schemes and anonymous credentials from bilinear maps, in: Advances in Cryptology - CRYPTO 2004, International Cryptologyconference, Santa Barbara, California, Usa, August 15-19, 2004, Proceedings, 2004, pp. 56–72.
[15] M. Stegelmann, D. Kesdogan, Gridpriv: a smart metering architecture offering k-anonymity, in: IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2012, pp. 419–426.
[16] M. Badra, S. Zeadally, Design and performance analysis of a virtual ring architecture for smart grid privacy, IEEE Trans. Inf. Forensics Secur. 9 (2) (2014) 321–329.
[17] P.J. Riesch, X. Du, Audit based privacy preservation for the openid authentication protocol, Homel. Secur. (2012) 348–352.
[18] X. Tan, J. Zheng, C. Zou, Y. Niu, Pseudonym-based privacy-preserving scheme for data collection, Smart Grid 22 (2) (2016) 120.
[19] Y. Sun, R. Lu, X. Lin, X. Shen, An efficient pseudonymous authentication scheme with strong privacy preservation for vehicular communications, IEEE Trans. Veh. Technol. 59 (7) (2010) 3589–3603.

[20] R. Lu, X. Lin, T.H. Luan, X. Liang, Pseudonym changing at social spots: an effective strategy for location privacy in vanets, IEEE Trans. Veh. Technol. 61 (1) (2012) 86–96.

[21] J. Yao, P. Venkitasubramaniam, The privacy analysis of battery control mechanisms in demand response: revealing state approach and rate distortion bounds, Decis. Control (2014) 1377–1382.

[22] G. Kalogridis, Z. Fan, S. Basutkar, Affordable privacy for home smart meters, in: Ninth IEEE International Symposium on Parallel and Distributed Processing with Applications Workshops, 2011, pp. 77–84.

[23] Z. Shi, R. Sun, R. Lu, L. Chen, Diverse grouping-based aggregation protocol with error detection for smart grid communications, IEEE Trans. Smart Grid 6 (6) (2015), 1–1.

[24] 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.

[25] F. Borges, M. Muhlhauser, Eppp4sms: efficient privacy-preserving protocol for smart metering systems and its simulation using real-world data, IEEE Trans. Smart Grid 5 (6) (2014) 2701–2708.

[26] E.J.G.D. Boneh, K. Nissim, Evaluating 2-dnf formulas on ciphertexts, in: Theory of Cryptography, Second Theory of Cryptography Conference, TCC 2005, Cambridge, MA, USA, February 10-12, 2005, Proceedings, 2005, pp. 325–341.

[27] L. Chen, R. Lu, Z. Cao, K. Alharbi, X. Lin, Muda: multifunctional data aggregation in privacy-preserving smart grid communications, Peer-to-Peer Netw. Appl. 8 (5) (2015) 1–16.

[28] Y. Gong, Y. Cai, Y. Guo, Y. Fang, A privacy-preserving scheme for incentive-based demand response in the smart grid, IEEE Trans. Smart Grid 7 (3) (2015), 1–1.

[29] H. Park, H. Kim, K. Chun, J. Lee, S. Lim, I. Yie, Untraceability of group signature schemes based on bilinear mapping and their improvement, in: International Conference on Information Technology, 2007, pp. 747–753.

[30] L.J. Pang, Y.M. Wang, A new (t,n) multi-secret sharing scheme based on shamirs secret sharing, Appl. Math. Comput. 167 (2) (2005) 840–848.

[31] A. Barletta, C. Callegari, S. Giordano, M. Pagano, Privacy preserving smart grid communications by verifiable secret key sharing, in: International Conference on Computing and Network Communications, 2015, pp. 199–204.

[32] A. Beussink, K. Akkaya, I.F. Senturk, M.M.E.A. Mahmoud, Preserving consumer privacy on ieee 802.11s-based smart grid ami networks using data obfuscation, in: 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS, 2014, pp. 658–663.

[33] C. Dwork, Differential Privacy, Springer Berlin Heidelberg, 2006.

[34] Y. Hong, J. Vaidya, H. Lu, P. Karras, S. Goel, Collaborative search log sanitization: toward differential privacy and boosted utility, Dependable & Secure Comput. IEEE Trans 12 (5) (2015) 504–518.