A Novel Privacy-Preserving Set Aggregation Scheme for Smart Grid Communications

Rongxing Lu[†], Khalid Alharbi[‡], Xiaodong Lin[‡], and Cheng Huang[†]
[†]School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
[‡]Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, Ontario, Canada
Email: rxlu@ntu.edu.sg; {khalid.alharbi, xiaodong.lin}@uoit.ca; huangcheng@ntu.edu.sg

***Abstract*—In this paper, we propose a novel privacy-preserving set aggregation scheme for smart grid communications. The proposed scheme is characterized by employing a group $\mathbb{G}$ of composite order $n = pq$ to achieve two-subset aggregation from a single aggregated data. With the proposed set aggregation scheme, the control center in smart grid is able to obtain more fine-grained data aggregation results for better monitoring and controlling smart grid. Detailed security analysis shows that the proposed scheme can achieve privacy-preserving property with formal proof in the random oracle model. In addition, extensive experiments are conducted, and the results demonstrate the proposed scheme is also efficient in terms of low computational costs and communication overheads.**

***Keywords* – Smart grid, Security, Privacy-preserving aggregation, Set aggregation**

I. Introduction

As the next generation of power grid, smart grid integrates various information and communication technologies (ICT) into power system to enable the power distribution more reliable and efficient from the power generation, transmission, distribution to end users, as shown in Fig. 1. Specifically, due to the two-way communications, smart grid can offer many benefits to both utilities and consumers [1], including i) overhauling aging equipments in current power system; ii) equipping the power grid to meet increasing demand; iii) decreasing brownouts, blackouts, and surges; iv) giving users to control over their power bills; v) facilitating real-time troubleshooting; vi) reduceing expenses to energy producers; and vii) making renewable power feasible.

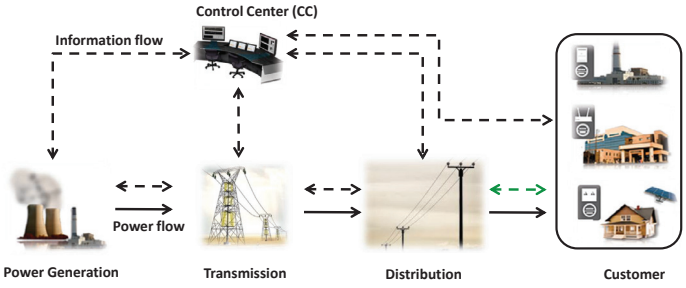


Fig. 1. Smart grid - the next generation of power grid

Smart meter is one of important components in smart grid communications, which enables residential users to report their nearly real-time electronic consumption data, e.g., every 15 minutes, to the control center for more reliable monitoring the health of power grid. However, the nearly real time reporting poses a potential threat to user privacy, e.g., personal information at home could be inferred by user's continuously reported data. Therefore, privacy-preserving technique must be employed in user data reporting for enhancing user's confidence in utilizing smart grid technique.

In recent years, to address the above challenge, many privacy-preserving data aggregation schemes have been proposed for smart grid communications [2]–[10]. However, most of them only support the data aggregation for the whole user set, which sometimes cannot meet the requirements from control center in smart grid communications. For example, the control center needs to know not only the total electronic consumption of the whole set of users, but also the number of users whose electronic consumption is higher than a threshold and the total consumption of these users. Although our previous EPPA work [10] can deal with this kind of set aggregation to some extent in smart grid, however, if the control center is assumed as honest-but-curious, each individual user's data can still be obtained by the control center. In this paper, in order to completely resolve the above problem, we propose a novel privacy-preserving set aggregation scheme for smart grid communications. Specifically, the main contributions of this paper are three-fold.

- Firstly, by using a group of composite order, we propose a novel privacy-preserving set aggregation scheme. Given a threshold of electronic consumption data, users can be divided into two subsets, then the proposed scheme can use one single aggregated data to aggregate the sum of electronic consumption data in each subset and the corresponding subset size in a privacy-preserving way, which thus supports more accurate data analytics for controlling and monitoring in smart grid.
- Secondly, with formal security proof technique, we show our proposed scheme can achieve each individual user's data privacy preservation.
- Finally, we implement our proposes scheme in Java and run extensive experiments to validate its efficiency in terms of low computational cost and communication overhead.

The remainder of this paper is organized as follows. In Section II, we formalize the system model, security model, and identify our research goal. We present the detailed design of our privacy-preserving set aggregation scheme in Section III, followed by the security analysis and performance evaluation in Section IV and Section V, respectively. Section VI reviews some related works and Section VII closes the paper with the conclusion.

II. Models and Design Goal

In this section, we formalize our system model, security model, and identify our design goal on set aggregation in smart grid communications.

A. System Model

In our system model, we focus on the set aggregation at the residential users in smart grid communications. In such a way, our system model mainly includes the following entities: a trusted authority (TA), a control center (CC), a residential gateway (GW) and a set of residential users $\mathbb{U} = \{U_1, U_2, ..., U_N\}$, as shown in Fig. 2, where N indicates the number of users in the set $\mathbb{U}$, and its maximal value is denoted as $N_{\max}$.

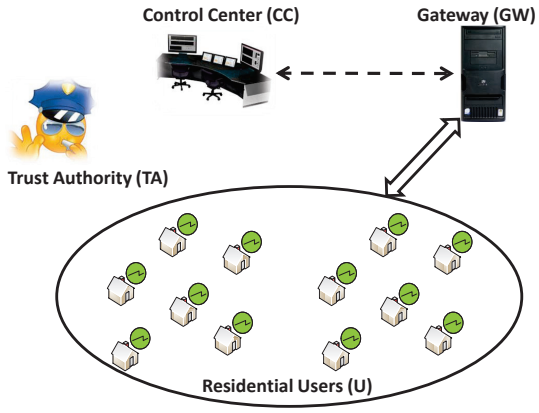


Fig. 2. System model under consideration

- Trust Authority (TA): TA is a fully trustable entity, whose duty is to manage and distribute key materials to other entities in the system. In general, after key distribution, TA will not be involved in the subsequent data aggregation process.
- Control Center (CC): CC is the core entity in the system, who is responsible for data collecting, processing and analyzing the nearly real-time data from $\mathbb{U}$ for monitoring the health of smart grid.
- Residential Gateway (GW): GW serves as a relay and aggregator role in the system, i.e., GW relays the information from CC to $\mathbb{U}$, and at the same time collects and aggregates the data from $\mathbb{U}$, and forwards the aggregated data to the CC.
- Residential Users $\mathbb{U} = \{U_1, U_2, ..., U_N\}$: Each user $U_i \in \mathbb{U}$ is equipped with smart meter, which collects and reports the nearly real-time electricity usage data m_i, e.g., every 15 minutes, to the CC via the GW.

Different from those previously reported data aggregation schemes [2]–[10], the *set aggregation* in smart grid communications enables the CC to obtain not only the whole aggregated result $\sum_{U_i \in \mathbb{U}} m_i$ for the set $\mathbb{U}$, but also the partial aggregated result $\sum_{U_j \in \mathbb{U}_1} m_j$ and the size $|\mathbb{U}_1|$ of a subset $\mathbb{U}_1 \subset \mathbb{U}$ from one single aggregated data, where $\mathbb{U}_1 = \{U_j | m_j \geq th\}$ and th is a user electronic consumption threshold. With this kind of set aggregation, the CC can make more accurate data analytics for monitoring and controlling the smart grid.

B. Security Model

In our security model, we consider both the CC and the GW are *honest-but-curious*. That is, they will faithfully follow the set aggregation protocol, but also attempt to get to know each individual user's data once the condition is satisfied. In addition, residential users $\mathbb{U} = \{U_1, U_2, ..., U_N\}$ are also honest, i.e., each U_i won't report false data the CC or collude with the CC to get other users' individual data.

Note that there are possible other attacks, i.e., bad data injection attack [11], DDoS attack, in smart grid communications. Since our focus is on privacy-preserving set aggregation, those attacks are currently beyond the scope of this work, and will be discussed in future work.

C. Design Goal

Our design goal is to develop an efficient and privacy-preserving set aggregation scheme for smart grid communication such that the CC can obtain more nearly real-time information from one single aggregated data. Specifically, the following two desirable goals should be satisfied.

- *The proposed scheme should be privacy-preserving.* Only the CC can read the set aggregation results in the proposed scheme, and no one (including the CC) can read each individual user data.
- *The proposed scheme should be efficient.* Not only the encryption at user side, aggregation at gateway, but also the decryption at control center should be efficient in terms of computational cost. In addition, the set aggregation, like other data aggregation schemes [2]–[10], should use one single aggregation data for transmission so as to achieve communication efficiency.

III. Proposed Privacy-Preserving Set Aggregation Scheme

In this section, we propose our privacy-preserving set aggregation scheme, which is mainly comprised of three parts: system initialization, encryption at user side, aggregation at gateway, and decryption at control center. Before plunging into the details, we first review hard problems in group with composite order [12], which serves as the basis of the proposed scheme.

A. Hard Problems in Group with Composite Order

Let κ be the security parameter and $\mathbb{G}(g, \times)$ be a cyclic multiplication group generated by g with composite order n, where $n = pq$ and p, q are two large prime numbers with $|p| =$

$|q| = \kappa$. Then, the Decisional Diffie-Hellman (DDH) Problem and Subgroup Decision (SD) Problem in $\mathbb{G}$ are described as follows:

Definition 1 (DDH Problem): Given $(g, g^a, g^b, Z) \in \mathbb{G}$ with unknown $a, b \in \mathbb{Z}_n^*$, to determine whether or not $Z = g^{ab} \in \mathbb{G}$. We say that (τ, ϵ)-DDH Assumption holds in $\mathbb{G}$ if no τ-time algorithm has advantage at least ϵ in solving the above DDH problem in $\mathbb{G}$.

Definition 2 (SD Problem): Without knowing the factorization of the group order $n = pq$, given an element $x \in \mathbb{G}$, to decide whether x is an element of subgroup in $\mathbb{G}$ with order p or q. We say that (t, ϵ)-SD Assumption holds in $\mathbb{G}$ if no t-time algorithm has advantage at least ϵ in solving the above SD problem in $\mathbb{G}$.

B. Description of The Proposed Scheme

1) System Initialization: Given the security parameter κ, a number $N_{\max}$ indicating the maximal user number in $\mathbb{U}$, a small number Δ indicating the maximal electricity consumption data in every t interval time, the control center (CC) first randomly chooses two large primes p, q such that $|p| = |q| = \kappa$, $2q > p$, and $p - q > N_{\max} \cdot \Delta$, computes $n = pq$, and also chooses a cyclic multiplication group $\mathbb{G}$ generated by g with the composite order n. After that, the CC chooses a cryptographic hash function $H : \{0,1\}^* \to \mathbb{G}$ and computes $h_0 = g^q$ and $h_1 = g^p$ in $\mathbb{G}$. Finally, the CC keeps $sk = p$ as the private key, and publishes the public key $pk = (n, g, h_0, h_1, H)$.

After verifying the validation of the CC's public key pk, the trusted authority (TA) chooses N random numbers $x_i \in \mathbb{Z}_n^*$, $i = 1, 2, \cdots, N$, and computes $x_0 \in \mathbb{Z}_n^*$ such that

$$x_0 + \sum_{i=1}^{N} x_i = 0 \bmod n \tag{1}$$

Finally, the TA sends x_0 as an additional decryption secret key to the CC, and x_i as a secret key to each corresponding user $U_i \in \mathbb{U} = \{U_1, U_2, \cdots, U_N\}$ via secure channels.

2) Encryption at User Side: At every time interval t, e.g., every 15 minutes, the CC chooses a threshold th and requests set aggregation from all residential users $\mathbb{U} = \{U_1, U_2, \cdots, U_N\}$. If a user U_i's electricity consumption data m_i is greater than or equal to the threshold th, i.e., $m_i \geq th$, U_i lies in the subset $\mathbb{U}_1 \subset \mathbb{U}$. Otherwise, U_i lies in the subset $\mathbb{U}_0 \subset \mathbb{U}$. Obviously, $\mathbb{U} = \mathbb{U}_1 \cup \mathbb{U}_0$, $\mathbb{U}_1 \cap \mathbb{U}_0 = \phi$.

Each user $U_i \in \mathbb{U}$ runs the following steps to encrypt his electricity consumption data m_i. Note that, because the time interval t is small, it is reasonable to assume m_i lies in a small set $\{0, 1, 2, \cdots, \Delta\}$.

- *Step 1:* U_i compares his consumption data m_i with the threshold th. If $m_i \geq th$, $U_i \in \mathbb{U}_1$ uses his secret key x_i to compute
$$c_i = g^{m_i} \cdot h_1 \cdot H(t)^{x_i} \tag{2}$$
Otherwise, if $m_i < th$, $U_i \in \mathbb{U}_0$ computes
$$c_i = h_0^{m_i} \cdot H(t)^{x_i} \tag{3}$$
- *Step 2:* U_i sends c_i to the gateway (GW).

3) Aggregation at Gateway: After receiving all c_i, $i = 1, 2, \cdots, N$, from the residential users $\mathbb{U}$, the GW performs the following aggregation,

$$\begin{aligned} C =& \prod_{i=1}^{N} c_i \\ =& g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{\sum_{U_i \in \mathbb{U}_1} 1} \cdot H(t)^{\sum_{i=1}^{N} x_i} \\ =& g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|} \cdot H(t)^{\sum_{i=1}^{N} x_i} \end{aligned} \tag{4}$$

and forwards the result C to the CC.

4) Decryption at Control Center: Upon receiving C, the CC performs the following steps to recover the aggregated data

- *Step 1:* the CC uses his secret key x_0 to compute
$$\begin{aligned} D =& C \cdot H(t)^{x_0} \\ =& g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|} \cdot H(t)^{\sum_{i=1}^{N} x_i + x_0} \\ & \xrightarrow{\because \sum_{i=1}^{N} x_i + x_0 = 0 \bmod n} \\ =& g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|} \end{aligned} \tag{5}$$
- *Step 2:* Because the CC knows h_0 is an element in the subgroup of $\mathbb{G}$ with order p, the CC uses the private key p to compute
$$\begin{aligned} \bar{D} =& D^p \\ =& \left(g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|} \right)^p \\ & \xrightarrow{\because h_0^p = 1} \\ =& \left(g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_1^{|\mathbb{U}_1|} \right)^p \\ =& h_1^{\sum_{U_i \in \mathbb{U}_1} m_i + p \cdot |\mathbb{U}_1|} \end{aligned} \tag{6}$$
and applies the Algorithm 1 to obtain the values of $\sum_{U_i \in \mathbb{U}_1} m_i$ and $|\mathbb{U}_1|$, where $|\mathbb{U}_1|$ is the size of subset $\mathbb{U}_1$.

Algorithm 1 Decrypt $\sum_{U_i \in \mathbb{U}_1} m_i$ and $|\mathbb{U}_1|$

```
procedure DECRYPTION
    on input of D̄ = h_1^{Σ_{U_i∈U_1} m_i + p·|U_1|}
    for (i = 0; i <= N; i++) do
        for (j = 0; j <= N·Δ; j++) do
            if (h_1^j · (h_1^p)^i == D̄) then
                set Σ_{U_i∈U_1} m_i = j, |U_1| = i
                break
            end if
        end for
    end for
    return Σ_{U_i∈U_1} m_i, |U_1|
end procedure
```

- *Step 3:* After calculating $\sum_{U_i \in \mathbb{U}_1} m_i$ and $|\mathbb{U}_1|$ in the last step, the CC now computes
$$\hat{D} = \frac{D}{g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_1^{|\mathbb{U}_1|}} = h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \tag{7}$$

Because $\sum_{U_j \in \mathbb{U}_0} m_j$ is in the range of $[0, N \cdot \Delta]$, $\sum_{U_j \in \mathbb{U}_0} m_j$ can be efficiently recovered from $\hat{D}$ by using Pollard's lambda method [13].

- *Step 4:* Because $\mathbb{U} = \mathbb{U}_1 \cup \mathbb{U}_0$, $\mathbb{U}_1 \cap \mathbb{U}_0 = \phi$, we can calculate the size of the subset $\mathbb{U}_0$ as $|\mathbb{U}_0| = N - |\mathbb{U}_1|$. Finally, the CC obtains the set aggregation shown in Table I. From the results in Table I, the CC can easily compute the whole aggregation value

$$\sum_{i=1}^{N} m_i = \sum_{U_i \in \mathbb{U}_1} m_i + \sum_{U_j \in \mathbb{U}_0} m_j \tag{8}$$

and run other more accurate data analytics algorithms for controlling and monitoring the smart grid.

TABLE I
THE RESULTS OF SET AGGREGATION

	Size	Aggregated Data
$\mathbb{U}_1$	$\lvert\mathbb{U}_1\rvert$	$\sum_{U_i \in \mathbb{U}_1} m_i$
$\mathbb{U}_0$	$\lvert\mathbb{U}_0\rvert$	$\sum_{U_j \in \mathbb{U}_0} m_j$

Correctness. Obviously, the correctness of the proposed scheme depends upon whether the Algorithm 1 can produce a unique solution (x, y) such that $\sum_{U_i \in \mathbb{U}_1} m_i = x$, $|\mathbb{U}_1| = y$, and $0 \le x \le N \cdot \Delta$, $0 \le y \le N$. In the following, we use Theorem 1 to show its correctness.

Theorem 1: Let $\hat{D}$ be $h_1^{\sum_{U_i \in \mathbb{U}_1} m_i + p \cdot |\mathbb{U}_1|}$ derived from a valid aggregation ciphertext C with the operations in Eqs. (5)-(6). Then, there exists a unique solution (x, y) such that $\sum_{U_i \in \mathbb{U}_1} m_i = x$, $|\mathbb{U}_1| = y$, and $0 \le x \le N \cdot \Delta$, $0 \le y \le N$.

Proof: Assume that, by running Algorithm 1, we have two solutions (x, y), (x', y') such that

$$\hat{D} = h_1^{x + p \cdot y} = h_1^{x' + p \cdot y'}$$

where $0 \le x, x' \le N \cdot \Delta$ and $0 \le y, y' \le N$. Because the order of h_1 is q, i.e., ${h_1}^q = 1$, we have

$$x + py = x' + py' \bmod q$$

Without loss of generality, we assume $y > y'$, then

$$\begin{aligned} x' =& x + p(y - y') \bmod q \\ =& [(x \bmod q) + (p \bmod q) \cdot ((y - y') \bmod q)] \bmod q \\ & \xrightarrow{\because 2q > p \text{ and } p - q > N_{\max} \cdot \Delta, \therefore (p \bmod q) > N_{\max} \cdot \Delta} \\ >& [(x \bmod q) + N_{\max} \cdot \Delta \cdot ((y - y') \bmod q)] \bmod q \\ =& x + N_{\max} \cdot \Delta \cdot (y - y') \bmod q \\ \ge& x + N \cdot \Delta \cdot (y - y') \bmod q \end{aligned}$$

which indicates that $x' > N \cdot \Delta$ and contradicts with the constraint $0 \le x' \le N \cdot \Delta$. Therefore, there only exists one unique solution (x, y) such that $\sum_{U_i \in \mathbb{U}_1} m_i = x$, $|\mathbb{U}_1| = y$, and $0 \le x \le N \cdot \Delta$, $0 \le y \le N$. ∎

Note that, as the three values of $\sum_{U_i \in \mathbb{U}_1} m_i$, $\sum_{U_j \in \mathbb{U}_0} m_j$, and $|\mathbb{U}_1|$ in $D = g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|}$ are small, it is possible to use the brute force method to directly guess a solution (x, y, z) such that $x \in [0, N \cdot \Delta], y \in [0, N \cdot \Delta], z \in [0, N]$ and $D = g^x {h_0}^y {h_1}^z$. However, the complexity is $O(N^3 \cdot \Delta^2)$, while the proposed scheme only requires $O(N^2 \cdot \Delta + \sqrt{N \cdot \Delta})$.

IV. Security Analysis

In this section, we analyze the privacy properties of the proposed scheme. In specific, following the security model discussed earlier, we will show that i) the CC cannot read each individual user's data, and ii) no one, except the CC, can read the set aggregation results.

A. The CC cannot read each individual user's data in the proposed scheme.

First, no matter whether a user U_i is in subset $\mathbb{U}_1$ or $\mathbb{U}_0$, we can always unify U_i's message M_i and the corresponding ciphertext c_i as

$$M_i = g^{a_{i1}} h_0^{a_{i2}} h_1^{a_{i3}}, \quad c_i = g^{a_{i1}} h_0^{a_{i2}} h_1^{a_{i3}} H(t)^{x_i} = M_i H(t)^{x_i}$$

where

$$\begin{cases} a_{i1} = m_i, a_{i2} = 0, a_{i3} = 1, & \text{if } U_i \in \mathbb{U}_1; \\ a_{i1} = 0, a_{i2} = m_i, a_{i3} = 0, & \text{if } U_i \in \mathbb{U}_0. \end{cases}$$

Based on the above decryption procedure at the control center, only if knowing $M_i = g^{a_{i1}} h_0^{a_{i2}} h_1^{a_{i3}}$, the CC can use the private key p to recover (a_{i1}, a_{i2}, a_{i3}). Therefore, in order to keep U_i's data privacy, we need to prove that the CC cannot get M_i from $c_i = M_i H(t)^{x_i}$. To formally show this point, we first assume U_i's public key $Y_i = g^{x_i}$ corresponding to the secret key x_i is available to the CC. Then, we prove in Theorem 2 that, even though the CC obtains Y_i, the CC still cannot know M_i under DDH assumption and in the random oracle model [14].

Theorem 2: Let $\mathcal{A}$ be any chosen-plaintext adversary against the user U_i's ciphertext $c_i = M_i \cdot H(t)^{x_i}$ with time τ. After q_h queries to the random oracles, its advantage is a non-negligible ϵ. Then, the DDH problem in $\mathbb{G}$ can be solved with another probability ϵ' with time τ', where

$$\epsilon' = \frac{\epsilon}{2}, \quad \tau' \le \tau + q_h \cdot T_h$$

with T_h denotes the time cost for each hash query.

Proof: We now use sequence games [15] to formally prove the theorem, i.e., showing the ciphertext $c_i = M_i \cdot H(t)^{x_i}$ is indistinguishable against $\mathcal{A}$ under chosen-plaintext attack (IND-CPA). We define a sequence of games $Game_1, Game_2, \cdots$ of modified attacks starting from the actual game $Game_0$. With these incremental games, we reduce a DDH problem instance, i.e., given (g, g^a, g^b, Z) for unknown $a, b \in \mathbb{Z}_n^*$ to determine whether or not $Z = g^{ab}$, to an IND-CPA attack against $c_i = M_i \cdot H(t)^{x_i}$. In other words, we will show that the adversary $\mathcal{A}$ can help to solve the DDH problem in $\mathbb{G}$.

$Game_0$: This is a real game in the random oracle model. We take the role of user U_i, and know the public and private key pair $(Y_i = g^{x_i}, x_i)$. The adversary $\mathcal{A}$ knows the public key $Y_i = g^{x_i}$ and is allowed to access a random oracle $\mathcal{O}_H$. At some time, $\mathcal{A}$ outputs two messages (M_{i0}, M_{i1}) and a time point t^* for encryption query. Then, we toss a coin to get a

random $\beta \in \{0,1\}$, encrypt and return $c_i^* = M_{i\beta} \cdot H(t^*)^{x_i}$ to $\mathcal{A}$. Finally, $\mathcal{A}$ outputs his guess $\beta' \in \{0,1\}$ on β. We denote $Guess_0$ as the event that $\beta = \beta'$ in $Game_0$ and use the notation $Guess_j$ for the same meaning in any game $Game_j$. Then, based on the definition, we have

$$\Pr[Guess_0] = \Pr[\beta = \beta'], \quad \epsilon = 2\Pr[\beta = \beta'] - 1$$

$Game_1$: In this game, we embed the challenge (g, g^a, g^b, Z) into the game, i.e., simulating the random oracle $\mathcal{O}_H$ by maintaining a hash list Λ_H. When a fresh query on time t_i is asked, we first choose a random number $r_i \in \mathbb{Z}_n^*$, set and return $H(t_i) = g^{b \cdot r_i}$ to $\mathcal{A}$, and also store $(t_i, r_i, H(t_i))$ in Λ_H. Because $H(t_i) = g^{b \cdot r_i}$ is uniformly distributed in $\mathbb{G}$, as a result this game is perfectly indistinguishable from the previous one. Therefore

$$\Pr[Guess_1] = \Pr[Guess_0]$$

$Game_2$: In this game, we replace the public key $Y_i = g^{x_i}$ with g^a. Once $Y_i = g^a$, we do not know the corresponding private key a. Therefore, when $\mathcal{A}$ sends (M_{i0}, M_{i1}, t^*) for a request, we perform the following simulation steps.

- Find the entry $(t_x, r_x, H(t_x))$ in Λ_H such that $t_x = t^*$.
- Compute $c_i = M_{ib^*} \cdot Z^{r_x}$ and return c_i to $\mathcal{A}$.

Now, we define the event B that $Z = g^{ab} \in \mathbb{G}$. If the event B really happens, then $c_i = M_{i\beta} \cdot Z^{r_x}$ is a valid ciphertext under the public key $Y_i = g^a$ and $H(t^*) = g^{b \cdot r_x}$. Therefore, the adversary $\mathcal{A}$ can exert his capability to guess whether $\beta = \beta'$ on $c_i = M_{i\beta} \cdot Z^{r_x}$. That is,

$$\Pr[Guess_2|B] = \Pr[Guess_1]$$

However, if the event B doesn't occur, i.e., $Z \neq g^{ab} \in \mathbb{G}$, then $\mathcal{A}$ can only randomly guess whether $\beta = \beta'$ with $1/2$ probability. Thus,

$$\Pr[Guess_2|\bar{B}] = \frac{1}{2}$$

Therefore, from the above analysis, we can solve the DDH problem in $\mathbb{G}$ with the advantage ϵ', where

$$\begin{aligned}\epsilon' &= \Pr[Guess_2|B] - \Pr[Guess_2|\bar{B}] = \Pr[Guess_1] - \frac{1}{2} \\ &= \Pr[Guess_0] - \frac{1}{2} = \Pr[\beta = \beta'] - \frac{1}{2} = \frac{\epsilon}{2}\end{aligned}$$

By a simple computation, we can obtain the claimed bound for $\tau' \leq \tau + q_h \cdot T_h$. Thus, the proof is completed. ■

From Theorem 2, we can see, even though the adversary $\mathcal{A}$ knows $Y = g^{x_i}$, the ciphertext $c_i = M_i \cdot H(t)^{x_i}$ is still secure. Therefore, we can conclude that the CC cannot read each individual user's data, and identify whether a user in $\mathbb{U}_1$ or $\mathbb{U}_0$ in the proposed scheme.

B. No one, except the CC, can read the set aggregation results in the proposed scheme.

In the proposed scheme, the aggregation ciphertext C is in the form of

$$C = g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|} \cdot H(t)^{\sum_{i=1}^{N} x_i} \in \mathbb{G}$$

Without knowing the secret key x_0 such that $x_0 + \sum_{i=1}^{n} = 0 \mod n$, $H(t)^{\sum_{i=1}^{N} x_i}$ cannot be removed from C. Therefore, only the CC can compute

$$D = C \cdot H(t)^{x_0} = g^{\sum_{U_i \in \mathbb{U}_1} m_i} \cdot h_0^{\sum_{U_j \in \mathbb{U}_0} m_j} \cdot h_1^{|\mathbb{U}_1|}$$

and read the set aggregation results $\sum_{U_i \in \mathbb{U}_1} m_i$, $\sum_{U_j \in \mathbb{U}_0} m_j$, and $|\mathbb{U}_1|$ in the proposed scheme.

From the above analysis, we can conclude that our proposed scheme is a secure and privacy-preserving set aggregation scheme for smart grid communications.

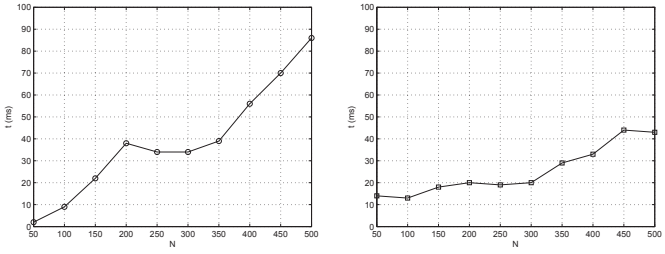
V. Performance Evaluation

In this section, we evaluate our proposed privacy-preserving set aggregation scheme in terms of computational cost and communications overheads. Specifically, we implement our scheme by Java (JDK 1.8) and run our experiments on a Laptop with 3.1 GHz processor, 8GB RAM, and Window 7 platform. The detailed parameter settings are shown in Table II.

TABLE II
The results of set aggregation

Parameter	Value
κ	$\kappa = 512$
$\mathbb{G}$	$\mathbb{G}$ is a subgroup of $\mathbb{Z}_P^*$ of order $n = pq$, where $P = 2pq+1$ is a large prime, and p, q are also two primes with $\|p\| = \|q\| = \kappa$
$N_{\max}$	$N_{\max} = 500$
N	$N = 50, 100, 150, 200, 250, 300, 350, 400, 450, 500$
Δ	$\Delta = 10$
th	the threshold th is randomly chosen from $[1, \Delta]$

Although the decryption complexity of the proposed scheme has been $O(N^2 \cdot \Delta + \sqrt{N\Delta})$, reduced from $O(N^3 \cdot \Delta^2)$, and can be acceptable by the powerful control center in smart grid communications, we still establish a hash table (stored in a zip file around 40 M) for accelerating the looking-up process in decryption in our experiment, where each entry in Hash table is the hash value of $h_1^j \cdot (h_1^p)^i$, with $0 \leq j \leq N_{\max} \cdot \Delta$ and $0 \leq i \leq N_{\max}$. We run our experiments 10 times, and the average results are reported below.



(a) Aggregation at GW (b) Decryption at CC

Fig. 3. Computational costs of aggregation and decryption varying with N

Computational cost. No matter whether a user belongs to subset $\mathbb{U}_1$ or $\mathbb{U}_0$, the average encryption at user side only takes 3.46 ms, which is extremely efficient. Fig. 3 shows the

computational costs of aggregation at GW and decryption at CC varies with the number of user N from 50 to 500 with the increment of 50. From the figure, we can see both of them are efficient, and the number of users N has a little effect on the aggregation and decryption, after a hash table used for looking-up in decryption is established in advance.

Communication cost. When $|p| = |q| = 512$, the length of $P = 2pq + 1$ is 1025 bits. Thus, any ciphertext (including c_i and C) in the subgroup $\mathbb{G}$ of $\mathbb{Z}_P^*$ is less than or equal to 1025 bits.

VI. Related Works

Recently, there are several privacy-preserving data aggregation schemes proposed for smart grid communications, we briefly review some of them [2]–[10] in this section.

Based on the Paillier homomorphic encryption, Garcia and Jacobs [2] first introduce a privacy-friendly energy-metering scheme for smart grid communications, and then Erkin and Tsudik [3] propose a more flexible privacy-preserving scheme supporting both spatial and temporal power consumption aggregation. Chen et al. [4] combine Paillier encryption and secret sharing technique to propose a fault-tolerant privacy-preserving data aggregation. In addition, based on some homomorphic techniques, other efficient schemes [5]–[9] have also been proposed, some of them are even secure against differential attack [6]–[8]. However, all above schemes are the whole user set aggregation, and cannot support partially subset aggregation. In [10], we propose an efficient and privacy-preserving aggregation scheme, called EPPA, which uses a super-increasing sequence to structure multi-dimensional data and encrypt the structured data with Paillier encryption. Inherently, EPPA can support set aggregation in smart grid communications. However, if the CC is honest-but curious in EPPA, the CC can read each individual user's data.

Different from the above schemes, our proposed scheme supports set aggregation in smart grid communications, thus it can provide more accurate information for smart grid monitoring. In addition, even though the CC is honest-but curious, the CC still cannot read each individual user's data.

VII. Conclusions

In this paper, we have proposed a novel privacy-preserving set aggregation scheme for smart grid communications. Given a threshold th of user electronic consumption data, the whole residential users $\mathbb{U}$ are divided into two subsets $\mathbb{U}_1$ and $\mathbb{U}_0$. The proposed scheme can just use one single aggregated ciphertext to aggregate the sum of electronic consumption data in each subset and the corresponding subset size, which thus supports more accurate data analytics for controlling and monitoring the smart grid. Detailed security analysis shows that the proposed scheme is privacy-preserving, i.e., no one can read each individual user's data, and only the CC can read the set aggregation results. Through extensive performance evaluation, we have also demonstrated that the proposed scheme is efficient in terms of computational costs and communication overhead. In future work, we are ready to exploit multi-subset aggregation, and take data integrity and differential privacy into consideration.

Availability

The Java implementation of the proposed privacy-preserving set aggregation scheme can be downloaded at `http://www.ntu.edu.sg/home/rxlu/project/`.

Acknowledgment

The authors would like to thank the support of Nanyang Technological University under Grant NTU-SUG (M4081196), MOE Tier 1 (M4011177) and AOARD-144029.

References

[1] "Consumer benefits," http://www.whatissmartgrid.org/smart-grid-101/consumer-benefits.

[2] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Security and Trust Management - 6th International Workshop, STM 2010, Athens, Greece, September 23-24, 2010, Revised Selected Papers*, 2010, pp. 226–238.

[3] Z. Erkin and G. Tsudik, "Private computation of spatial and temporal power consumption with smart meters," in *Applied Cryptography and Network Security - 10th International Conference, ACNS 2012, Singapore, June 26-29, 2012. Proceedings*, 2012, pp. 561–577.

[4] 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 (PPNA) (Springer)*, to appear.

[5] K. Alharbi and X. Lin, "LPDA: A lightweight privacy-preserving data aggregation scheme for smart grid," in *International Conference on Wireless Communications and Signal Processing, WCSP 2012, Huangshan, China, October 25-27, 2012*, 2012, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/WCSP.2012.6542936

[6] E. Shi, T. H. Chan, E. G. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Proceedings of the Network and Distributed System Security Symposium, NDSS 2011, San Diego, California, USA, 6th February - 9th February 2011*, 2011.

[7] H. Bao and R. Lu, "A new differentially private data aggregation with fault tolerance for smart grid communications," *IEEE Internet of Things Journal*, to appear.

[8] L. Chen, R. Lu, Z. Cao, K. AlHarbi, and X. Lin, "MuDA: Multifunctional data aggregation in privacy-preserving smart grid communications," *Peer-to-Peer Networking and Applications (PPNA) (Springer)*, to appear.

[9] C. Li, R. Lu, H. Li, L. Chen, and J. Chen, "PDA: A privacy-preserving dual-functional aggregation scheme for smart grid communications," *Security and Communication Networks*, to appear.

[10] 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 Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1621–1631, 2012. [Online]. Available: http://dx.doi.org/10.1109/TPDS.2012.86

[11] Y. Liu, M. K. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 2009 ACM Conference on Computer and Communications Security, CCS 2009, Chicago, Illinois, USA, November 9-13, 2009*, 2009, pp. 21–32. [Online]. Available: http://doi.acm.org/10.1145/1653662.1653666

[12] D. Boneh, E. Goh, and 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.

[13] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 1996.

[14] M. Bellare and P. Rogaway, "Random oracles are practical: A paradigm for designing efficient protocols," in *CCS '93, Proceedings of the 1st ACM Conference on Computer and Communications Security, Fairfax, Virginia, USA, November 3-5, 1993.*, 1993, pp. 62–73. [Online]. Available: http://doi.acm.org/10.1145/168588.168596

[15] R. Lu, X. Lin, Z. Cao, J. Shao, and X. Liang, "New (t, n) threshold directed signature scheme with provable security," *Inf. Sci.*, vol. 178, no. 3, pp. 756–765, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2007.07.025