Front.Comput.Sci.
DOI

Supplementary File

A novel privacy preserving data aggregation scheme with data integrity and fault tolerance for smart grid communications

Haiyong Bao [1,2], **Beibei Li**(✉)[3]

1 Software Engineering Institute, East China Normal University, Shanghai 200062, China
2 School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
3 College of Cybersecurity, Sichuan University, Chengdu 610065, China

Abstract To design a secure and efficient data aggregation scheme satisfying real applications has been pursued by research communities for a long time. In this paper, we propose a novel secure data aggregation scheme to achieve privacy preservation, data integrity, and fault tolerance simultaneously for smart grid communications. Specifically, firstly, a hierarchical communication architecture of "RU-GW-CC" is constructed, and the system model, attack model, security and performance requirements are analyzed and formalized. Based on which, a concrete secure and efficient data aggregation scheme is instantiated. Furthermore, utilizing the *pre-cached auxiliary information*, a new fault tolerant mechanism for reliable smart grid communications is designed, which is able to aggregate measurements of normal smart meters efficiently and flexibly for any rational number of malfunctioning smart meters with arbitrary long failure period. In addition, efficient authentication techniques are studied to flexibly generate and share session keys in a non-interactive way, which is utilized for symmetric encryption algorithm to achieve data integrity and source authentication of communications. Finally, extensive performance evaluations are conducted to illustrate that the proposed data aggregation scheme outperforms the state-of-the-art similar schemes in terms of computation complexity, and communication cost.

Keywords Smart grid, Data aggregation, Privacy preservation, Data integrity, Fault tolerance.

Received month dd, yyyy; accepted month dd, yyyy

E-mail: *libeibei@scu.edu.cn*

1 Introduction

Smart grid, regarded as the next generation of power grid, has attracted wide attention in recent years [1–3]. By making the best of the information technology in smart grid, considerable power energy can be effectively saved. Specifically, as shown in Fig. 1, due to *information and communications technology* (ICT), together with smart grid's cyber-physical architecture, huge amount of information is collected and delivered to the control center (CC) for real-time monitoring and analyzing the health of power grid [4–6].

However, user's real-time data, *e.g.*, collected every 15 minutes, contain detailed power usage information, which is high relevant to user's lifestyle. Thus, it is of vital importance to preserve user's privacy. In addition to privacy preservation, data integrity is also critical in smart grid communications. Otherwise, an attacker could eavesdrop, intercept, and/or maliciously analyze energy usage data to reduce the availability of smart grid. Therefore, it is of great significance to preserve data privacy and ensure data integrity simultaneously in smart grid communications.

In order to address privacy issues, several privacy-preserving data aggregation schemes for smart grid communications have been proposed [1,7–12]. Most of them [1,7–9] utilize the homomorphic encryption techniques [13] to encrypt and aggregate data in the local area gateway (GW) and forward to CC. However, the private key CC held may be

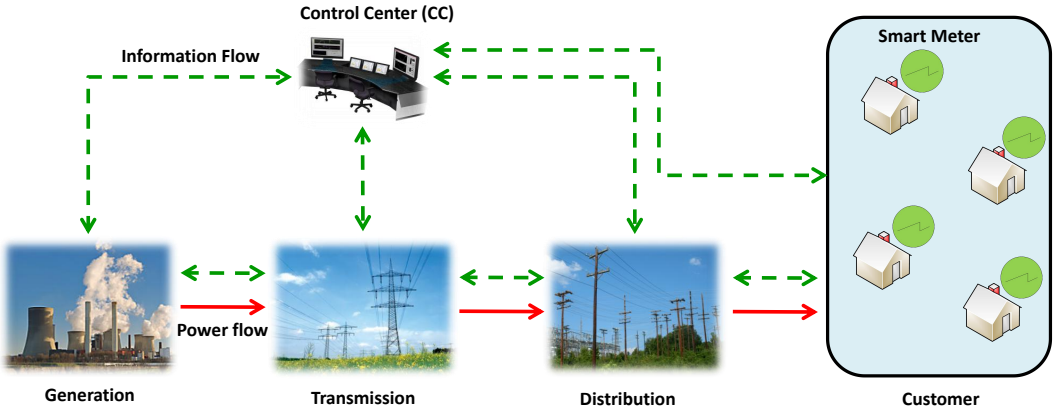


Fig. 1 Conceptual architecture of smart grid

abused not just to decrypt the aggregated information, but to reveal single user's electricity usage. More seriously, some strong adversaries may deploy undetectable malwares to G-W or CC for disclosure of user's privacy [7, 9]. When strong adversaries are considered, the above privacy-preserving data aggregation schemes are not robust enough, and conflict user's privacy concerns greatly. Other data aggregation schemes, *e.g.*, [8, 10, 11], introduce *blind factor embedding* technology, *i.e.*, the sum of all participants' random numbers equals to 0, to diminish CC's authority. Even if the single user fails to report data properly, CC will not be able to decrypt and obtain aggregated measurements successfully, because the sum of all embedded blind factors is not 0 any longer. On this account, one of the major drawbacks of such mechanisms is incapable of fault tolerance [8, 10]. Most existing data aggregation schemes do not support fault tolerance, and only a small part of them [11, 12] partially support it either with low availability or poor performance. Therefore, such schemes are not practical enough, especially when the exact number of fault smart meters (sometimes could be very large) can not be predicted. In addition to data aggregation, other anonymity techniques, *e.g.*, group signature, ring signature, *etc.*, have been studied to protect user's privacy [14, 15]. However, time-consuming and complex cryptography operations are often unavoidable in public key based anonymity techniques, which hinders the real application.

Meanwhile, in order to prevent adversaries from compromising user's report maliciously, *e.g.*, replay, inject, modify, forge, and/or delay, *etc.*, several message authentication mechanisms [16–18] have been designed to achieve data integrity in smart grid communications. Generally, the existing techniques mainly include Bins and Balls (BiBa) [17], Hash to Obtain Random Subsets Extension (HORSE) [16], Digital Signature Algorithm (DSA) [18], and Hash-based Message Authentication Code (MAC/HMAC) [19]. Even though DSA is more secure then BiBa and HORSE, it is at the cost of additional computational complexity, particularly at the receiver-end [20–22]. It is illustrated that MAC/HMAC based authentication technique is more efficient than DSA [19]. However, each round of public key-based session key agreement is essential for the integrity check of MAC/HMAC. It still brings high computing cost and communication overhead.

In addition, in the delay-sensitive and bandwidth-intensive smart grid communications, especially in the user side, the efficient mechanism with low computing cost and communication overhead should be designed to improve practicality.

Therefore, how to achieve an efficient, secure (with properties of privacy preservation and data integrity simultaneously), and reliable (with property of fault tolerance) data aggregation scheme for smart grid communications still deserves further investigations. In this paper, we propose a novel privacy preserving data aggregation scheme with data integrity and fault tolerance for smart grid communications. Specifically, the main contributions of this paper are three-fold.

- Firstly, based on the *pre-calculated auxiliary information*, a novel fault tolerant mechanism for reliable smart grid communications is designed, which is different from the traditional fault-tolerant method used trusted third party to track and distinguish normal and fault nodes. Specifically, utilizing the *pre-cached auxiliary information*, CC can obtain the aggregation of the functioning smart meters flexibly and efficiently for any rational number of malfunctioning smart meters with arbitrary long failure period.
- Secondly, by integrating a pair of identities and private/public keys of two communication parties, and dynamic reporting time point, a novel efficient authentica-

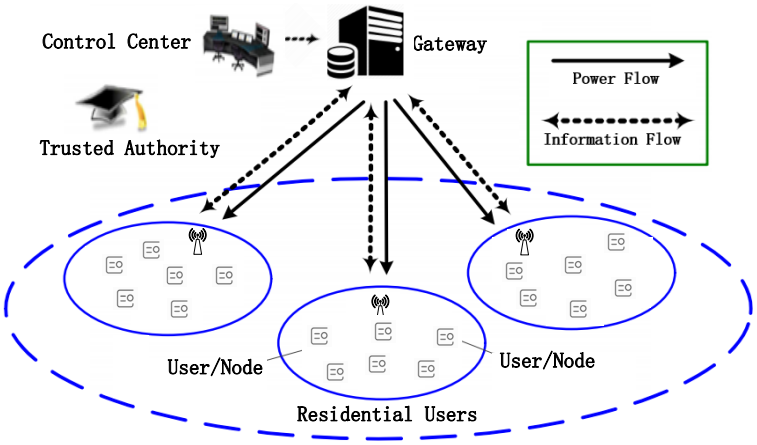


Fig. 2 System model under consideration

tion technique is proposed to flexibly generate and share session keys in a non-interactive way. The shared session key is utilized for symmetric encryption algorithm to achieve data integrity and source authentication of transmitted data. The security analysis indicates that the proposed scheme can efficiently and effectively prevent the malicious adversary from impairing (*e.g.*, replay, inject, modify, forge, and/or delay, *etc.*) the transmitted data.

- Thirdly, by constructing a hierarchical communication architecture of "RU-GW-CC", the proposed data aggregation system is logically decomposed and instantiated, which greatly improves the data exchange efficiency and protects the user's privacy simultaneously. In addition, through comparative performance analysis, we demonstrate that our proposed data aggregation scheme outperforms the state-of-the-art similar schemes in terms of computation complexity and communication cost.

The remainder of this paper is organized as follows. We first specify the problem formalization, including system model, attack model, and design goal in Section 2, and briefly review some preliminaries in Section 3. Then, our novel data aggregation scheme is presented in Section 4. Subsequently, the security analysis and performance evaluation are demonstrated in Section 5 and Section 6, respectively. We also discuss related works in Section 7. Finally, our conclusions are drawn in Section 8.

2 Problem formalization

In this section, we formalize the system model, attack model, and identify the design goal.

2.1 System model

Based on the typical scenario of smart grid communications, the overall system architecture is shown in Fig. 2, which includes the following four participants:

Trusted authority (TA), who is responsible for the management and distribution of other entities' secret information, and has high credibility and super computing power;

Control Center (CC), who is responsible for integrating, processing and analyzing the periodic time-series data from the residential users, and provides comprehensive, reliable, and intelligent services;

Gateway (GW), whose role is to connect CC and the residential users, and is responsible for aggregating data submitted by users and forwarding communications between residential users and CC;

Residential Users (RU), which includes n smart meters (users or nodes) in residential area (RA), *i.e.*, $U = \{U_1, \cdots, U_n\}$, who are responsible for real-time data collection and forwarding to CC through GW.

2.2 Attack model

In our attack model, GW and CC are considered trustworthy, and residential users $U = \{U_1, \cdots, U_n\}$ are also "honest" to abide the protocol. However, there is an external adversary $\mathcal{A}$, who may sneak into RA to intercept and maliciously analyze the communication packages, or invade servers of GW and CC to reveal user's privacy. In addition, $\mathcal{A}$ could launch some active attacks to destroy data integrity as well. Specifically, we consider the following frequent and common attacks in smart grid communications, which aims to reveal user's privacy and impair data integrity.

- *Privacy leakage attack*, where an adversary $\mathcal{A}$ may try

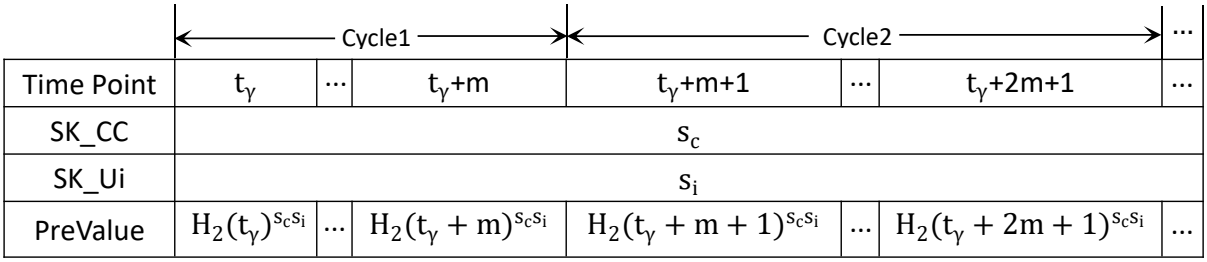


Fig. 3 Compute and cache fault-tolerant *auxiliary information* periodically and previously

to compromise the privacy of the residential user by eavesdropping, intercepting, and/or maliciously analyzing the communications from RA to GW and those from GW to CC;

- *Malware attack*, where an adversary $\mathcal{A}$ may deploy undetectable malwares to GW or CC for privacy disclosure of the residential user.
- *Data damage attack*, where an adversary $\mathcal{A}$ may intercept the communication packets and impair (replay, inject, modify, forge, and/or delay, *etc.*) the user's real data intended to report.

In addition to the above attacks, due to the physical malfunction, instrument aging, *etc.*, some smart meters could be malfunctioning or in failure status. The abnormal states giving rise to the relevant users to fail to report data normally, are also considered in our attack model.

It should be noted that our focus is to prevent external attacks from revealing user's privacy and/or disrupting the data integrity of communications. Other attacks, such as internal attacks, are beyond the scope of this study.

2.3 Design goal

Considering the above system model and attack model, our design goal is to propose a novel privacy preserving data aggregation scheme with data integrity and fault tolerance for smart grid communications. Specifically, the following design goals should be achieved.

- *Privacy preservation:* An external attacker $\mathcal{A}$ could not disclose user's privacy even though communication packages have been intercepted successfully. Meanwhile, although $\mathcal{A}$ could intrude into servers of GW or CC, and deploy some undetectable malwares, it still unable to reveal user's privacy either.
- *Data integrity:* The valid communications cannot be damaged during the transmission. If $\mathcal{A}$ replayed, injected, modified, forged, and/or delayed the communication packages, the malicious behaviors should be detected.
- *Fault tolerance:* Even in case of failure, the system should still be effective and efficient for data aggregation of malfunctioning smart meters.
- *Computation efficiency:* The proposed protocol should be computationally efficient to support data aggregation for thousands and millions of residential users.

3 Preliminary

In this section, we briefly recall the basic knowledge of Boneh-Goh-Nissim (BGN) cryptosystem [23] for the construction of our proposed scheme. The BGN cryptosystem is a public key encryption scheme. It has been widely used in many privacy-preserving applications since it can achieve some nice homomorphic properties. Specifically, the BGN cryptosystem includes three algorithms, *i.e.*, key generation, encryption, and decryption as follows.

- Key Generation: Given $\tau \in Z^+$, the security parameter, perform $\zeta(\tau)$ to get the tuple (p, q, G), where p, q are different primes such that $|p| = |q| = \tau$, and G is a cyclic group of order $N = pq$. Randomly chose two generators $g, x \in G$, and set $h = x^q$. Then, h is a random generator of the subgroup of G having order p. Eventually, the public key and the private key are $PK = (N, G, g, h)$ and $SK = p$, respectively.
- Encryption: Given a message $m \in \{0, 1, \cdots, V\}$, where $V << q$ is the bound of the message space, choose a random number $r \in Z_N$. Then, the ciphertext can be calculated as $C = g^m h^r \in G$.
- Decryption: Given the private key SK $= p$ and the ciphertext $C \in G$, first compute $C^p = (g^m h^r)^p = (g^p)^m$. Let $g_p = g^p$, then $C^p = g_p^m$. To recover m, it comes down to compute the discrete logarithm of g_p^m.

Note that when m is a short message, *i.e.*, $m \leq V$ for some small bound V, the decryption takes expected time $O(\sqrt{V})$ utilizing Pollard's lambda method [24].

4 Proposed Scheme

In this section, we propose our novel privacy preserving data aggregation scheme with data integrity and fault tolerance for smart grid communications. It mainly consists of six phases illustrated as follows.

4.1 System initialization

TA performs the following procedures to initialize the system:

- Based on the input of the secure parameter ρ, run $\zeta(\rho)$, and output the system parameters (G, g, p), where p is a safe prime, G is a cyclic group of order p, and the discrete logarithm problem in group G is difficult (*i.e.*, infeasible in computation). And randomly select g ∈ G, the generator of the group G.
- Perform the following procedures to assign the secret information of all users U ={$U_1, \cdots, U_n$}, GW and CC.
  - Randomly select n number of $s_i \in Z_p^*$, where i= 1, 2, . . . , n, and calculate S_i=g^{s_i}. The values of s_i and S_i are assigned as the private and public key of U_i (with identity information ID_i), respectively.
  - Compute $s_c \in Z_p^*$, satisfying $s_c \cdot (s_1 + \cdots + s_n) = 1 \ mod\ p$, and S_c=g^{s_c}. The values of s_c and S_c are assigned as the private and public key of CC (with identity information ID_c), respectively.
  - Randomly select $s_g \in Z_p^*$, and compute S_g=g^{s_g}. The values of s_g and S_g are assigned as the private and public key of GW (with identity information ID_g), respectively.
- Randomly select two hash functions: $H_1 : \{0,1\}^* \rightarrow G$ and $H_2 : \{0,1\}^* \rightarrow G$.
- Publish the system parameters: (G, p,g,S_g, S_c, ID_g, ID_c, H_1, H_2), and $< ID_i, S_i >$, where i= 1, 2, . . . , n.
- Select symmetric encryption algorithm AES. Denote AES_ENC_k and AES_DEC_k as the encryption algorithm, and decryption algorithm under the symmetric key k, respectively.

4.2 Data aggregation request

At m intervals of reporting time points, TA performs the following procedures, as shown in Fig. 3, to previously compute and cache the *auxiliary information* to support fault tolerance.

- Determine the current and future m number of reporting time points t_τ, where $\tau = t_\gamma, t_\gamma + 1, \ldots, t_\gamma + m$.
- Compute and cache $Y_{\tau,i} = H_2(t_\tau)^{s_c s_i} = h_\tau^{s_c s_i}$, where $\tau = t_\gamma, t_\gamma+1, \ldots, t_\gamma+m$ are the time dimension subscripts, and $i = 1, 2 \ldots, n$ are the user dimension subscripts. Because the planned reporting time points are known beforehand, the values of $Y_{\tau,i}$ can be pre-computed periodically.

At the current reporting time point t_τ, CC performs the following procedures to initiate the data aggregation requests.

- Compute h_τ=$H_2(t_\tau)$.
- Randomly select r ∈ Z_p^*, and calculate A_1=$h_\tau^{rs_c}$.
- Send A_1 to GW.

4.3 Data aggregation request relay

After receives A_1, GW forwards it to each user U_i, where i= 1, 2, . . . , n.

4.4 User data reporting

At the current reporting time point t_τ, each user U_i, where i=1, 2, · · · , n, performs the following procedures to report the current measurement $m_i \in \{0, 1, \cdots, W\}$ to GW.

- Compute g_τ=$H_1(t_\tau)$.
- Compute C_i=$g_\tau^{m_i} {A_1}^{s_i}$=$g_\tau^{m_i} h_\tau^{rs_i s_c}$.
- Compute the session key shared with GW in a non-interactive way as k_{ig}=$H_1(S_g{}^{s_i}||ID_i||ID_g||g_\tau)$=$H_1(g^{s_g s_i}||ID_i||ID_g||H_1(t_\tau))$.
- By performing the AES encryption algorithm, obtain the ciphertext $C_i' = AES\_ENC_{k_{ig}}(C_i||ID_i||ID_g||g_\tau) = AES\_ENC_{k_{ig}}(g_\tau^{m_i} h_\tau^{rs_i s_c}||ID_i||ID_g||H_1(t_\tau))$.
- Send $< C_i', ID_i >$ to GW.

4.5 Secure data aggregation

At the current reporting time point t_τ, GW computes $g_\tau = H_1(t_\tau)$, and performs the following procedures:
[1] *All users report data normally:*

- For each received $< C_i^{'}, ID_i >$, where i= 1, 2, . . . , n, according to ID_i, compute the corresponding session key shared with U_i in a non-interactive way as k_{ig}=$H_1(S_i{}^{s_g}||ID_i||ID_g||H_1(t_\tau))$=$H_1(g^{s_g s_i}||ID_i||ID_g||g_\tau)$.
- By performing the AES decryption algorithm, obtain the plaintext of each user U_i, where i= 1, 2, . . . , n, as $AES\_DEC_{k_{ig}}(C_i^{'})$ =$g_\tau^{m_i} h_\tau^{rs_i s_c}||ID_i||ID_g||H_1(t_\tau)$ = $C_i||ID_i||ID_g||g_\tau$.
- Compute the aggregated data of all users in RA as $C_g = \prod_{i=1}^{n} C_i$=$g_\tau^{\sum_{i=1}^{n} m_i} h_\tau^{rs_c \sum_{i=1}^{n} s_i}$=$g_\tau^{\sum_{i=1}^{n} m_i} h_\tau^{r}$.

- Compute the session key shared with CC in a non-interactive way as $k_{gc}=H_1(S_c{}^{s_g}||ID_g||ID_c||H_1(t_\tau))=H_1(g^{s_g s_c}||ID_g||ID_c||g_\tau)$.
- By performing the AES encryption algorithm, obtain the ciphertext as $C_g{}'=\text{AES\_ENC}_{k_{gc}}(C_g||ID_g||ID_c||H_1(t_\tau))=\text{AES\_ENC}_{k_{gc}}(g_\tau{}^{\sum_{i=1}^{n} m_i} h_\tau{}^{r}||ID_g\ ||ID_c||g_\tau)$.
- Send $C_g{}'$ to CC.

[2] *Some users do not report data normally:*

- Compute the session key shared with each user $U_i \in \widetilde{U}$, where $\widetilde{U}$ is the set of fault users, in a non-interactive way as $k_{ig}=H_1(S_i{}^{s_g}||ID_i||ID_g||H_1(t_\tau))=H_1(g^{s_g s_i}||ID_i||ID_g||g_\tau)$.
- By performing the AES decryption algorithm, obtain the plaintext of each user $U_i \in U/\tilde{U}$ as $\text{AES\_DEC}_{k_{ig}}(C_i{}')=g_\tau{}^{m_i} h_\tau{}^{rs_i s_c}||ID_i||ID_g||H_1(t_\tau)=C_i||ID_i||ID_g||g_\tau$.
- Compute the aggregated data of all users $U_i \in U/\tilde{U}$ who reported data successfully as $\tilde{C}_g=\prod_{U_i\in U/\tilde{U}} C_i=g_\tau{}^{\sum_{U_i\in U/\tilde{U}} m_i} h_\tau{}^{rs_c \sum_{U_i\in U/\tilde{U}} s_i}$.
- Compute the session key shared with CC in a non-interactive way as $k_{gc}=H_1(S_c{}^{s_g}||ID_g||ID_c||H_1(t_\tau))=H_1(g^{s_g s_c}||ID_g||ID_c||g_\tau)$.
- By performing the AES encryption algorithm, obtain the ciphertext as $\tilde{C}_g{}'=\text{AES\_ENC}_{k_{gc}}(\tilde{C}_g||ID_g||ID_c||H_1(t_\tau))=\text{AES\_ENC}_{k_{gc}}(g_\tau{}^{\sum_{U_i\in U/\tilde{U}} m_i}\ h_\tau{}^{rs_c \sum_{U_i\in U/\tilde{U}} s_i}||ID_g||ID_c||g_\tau)$.
- Send $\tilde{C}_g{}'$ to CC.

4.6 Aggregated data recovery

At the current reporting time point t_τ, CC computes $g_\tau = H_1(t_\tau)$ and $h_\tau=H_2(t_\tau)$, and performs the following procedures:
[1] *All users report data normally:*

- Compute the session key shared with GW in a non-interactive way as $k_{gc} = H_1(S_g{}^{s_c}||ID_g||ID_c||H_1(t_\tau))=H_1(g^{s_g s_c}||ID_g||ID_c||g_\tau)$.
- By performing the AES decryption algorithm, obtain the plaintext as $\text{AES\_DEC}_{k_{gc}}(C_g{}') = C_g||ID_g||ID_c||g_\tau = g_\tau{}^{\sum_{i=1}^{n} m_i} h_\tau{}^{r}||ID_g||ID_c||H_1(t_\tau)$.
- Compute $C_g h_\tau{}^{-r} = g_\tau{}^{\sum_{i=1}^{n} m_i}$.
- By computing the discrete log of $g_\tau{}^{\sum_{i=1}^{n} m_i}$, recover the aggregated data $\sum_{i=1}^{n} m_i$ in the expected time $O(\sqrt{nW})$ using Pollard's lambda method [24].

[2] *Some users do not report data normally:*

- CC sends $U_i \in \tilde{U}$, the ID set of the fault nodes, to TA. According to $\tilde{U}$, TA computes $\overline{C}_\tau=\prod_{U_i\in\widetilde{U}} Y_{t_\tau,i}=h_\tau{}^{s_c \sum_{U_i\in\widetilde{U}} s_i}$, and sends $\overline{C}_\tau$ to CC.
- CC computes the session key shared with GW in a non-interactive way as $k_{gc} = H_1(S_g{}^{s_c}||ID_g||ID_c||H_1(t_\tau))=H_1(g^{s_g s_c}||ID_g||ID_c||g_\tau)$.
- By performing the AES decryption algorithm, CC obtains the plaintext as $\text{AES\_DEC}_{k_{gc}}(\tilde{C}_g{}') = \tilde{C}_g||ID_g||ID_c||g_\tau = g_\tau{}^{\sum_{U_i\in U/\tilde{U}} m_i} h_\tau{}^{rs_c \sum_{U_i\in U/\tilde{U}} s_i}||ID_g||ID_c||H_1(t_\tau)$.
- CC computes $\tilde{C}_g \cdot (\overline{C}_\tau{}^{r}) \cdot (h_\tau^{-r}) = (\prod_{U_i\in U/\tilde{U}} C_i) \cdot (\prod_{U_i\in\tilde{U}} Y_{t_\tau,i})^r \cdot (h_\tau{}^{-r}) = g_\tau{}^{\sum_{U_i\in U/\tilde{U}} m_i} h_\tau{}^{rs_c \sum_{U_i\in U/\tilde{U}} s_i} h_\tau{}^{rs_c \sum_{U_i\in\tilde{U}} s_i} h_\tau{}^{-r} = g_\tau{}^{\sum_{U_i\in U/\tilde{U}} m_i}$.
- Similar as the corresponding procedures that all users report data normally, the aggregated data $\sum_{U_i\in U/\tilde{U}} m_i$ can be recovered successfully.

5 Security analysis

In this section, we will illustrate that our proposed data aggregation scheme achieves all the security requirements defined in Section 2.

- *The user's communication link is protected from privacy leakage attack.*
  Firstly, an adversary $\mathcal{A}$ may reside in RA to eavesdrop the communications. Suppose $\mathcal{A}$ has eavesdropped the report from U_i to GW at time point t_γ as $< C_i{}', ID_i >$. Because the user's measurement is encrypted by AES encryption algorithm as $C_i{}' = \text{AES\_ENC}_{k_{ig}}(C_i||ID_i||ID_g||g_\tau) = \text{AES\_ENC}_{k_{ig}}(g_\tau{}^{m_i} h_\tau{}^{rs_i s_c}||ID_i||ID_g||H_1(t_\tau))$, $\mathcal{A}$ cannot obtain the corresponding plaintext, provided that the session key k_{ig} for AES encryption, which is generated cooperatively by the mutual communication parties' private keys, is secure against $\mathcal{A}$. In the following, we will illustrate that even though the session key k_{ig} for some reporting time point, is exposed, $\mathcal{A}$ still cannot obtain the user's measurement m_i from AES plaintexts of $C_i=g_\tau{}^{m_i} A_1{}^{s_i}=g_\tau{}^{m_i} h_\tau{}^{rs_i s_c}$. Observing the electricity usage m_i within 15 minute is commonly a small value, $\mathcal{A}$ may try to launch the *brute-force attack* by exhaustedly testing each possible value of m_i. However, due to the discrete logarithm problem (DLP), $\mathcal{A}$ is unable to obtain the private keys of U_i and CC (*i.e.*, s_i and s_c), and the value of $r \in Z_p^*$, which was randomly selected by CC. Hence, the individual usage data m_i of U_i cannot be recovered.

Similarly, the communications from GW to CC are of the same form as U_i's report to GW. Thus, $\mathcal{A}$ cannot obtain the individual user's usage data via eavesdropping the communications.

When some smart meters, say $\hat{U} \subset U$, are malfunctioning, because the values of $\overline{\mathrm{C}}_\tau$ and r are kept secret by CC, without them, anyone else cannot recover the sum usage data of functioning smart meter $\sum_{U_i \in U/\tilde{U}} m_i$, let alone each user's private usage data m_i, even if the session key k_{gc} for AES encryption between GW and CC is exposed to $\mathcal{A}$, and $\mathcal{A}$ could obtain $\tilde{C}_g$, the corresponding AES plaintext of some reporting time point.

- *The user's communication link is protected from malware attack.*

  Even though $\mathcal{A}$, after deploying some undetectable malwares into GW or intruding into the database of GW, has stolen the stored data successfully, $\mathcal{A}$ could only get the aggregations and ciphertexts of all users' data. Because GW never decrypts any user's electricity usage data, $\mathcal{A}$ still cannot get any user's individual usage data. In addition, $\mathcal{A}$ could also intrude into the database of CC. However, after decryption, the outputs CC generated are still the aggregated ones, from which the single user's data is not revealed at all. Therefore, the individual user's report is protected from malwares attack.
- *The user's communication link is protected from data damage attack.*

  We will show that the user's communications cannot be altered during the transmissions.
  - Communication pollution attack resistance

    Firstly, we consider the communications from U_i to GW. Upon receiving $< C_i', ID_i >$, according to ID_i, GW first computes the non-interactive session key k_{ig}=$H_1(S_i{}^{s_g}||ID_i||ID_g||H_1(t_\tau))$=$H_1(g^{s_g s_i}||ID_i||ID_g||g_\tau)$ shared with U_i. Then, GW performs the AES decryption using k_{ig} to obtain U_i's report. Because the secret keys of s_i and s_g are utilized to compute the shared session key k_{ig} collaboratively by U_i and GW, $\mathcal{A}$ cannot obtain the agreed secret key, nor can it alter the original data encrypted by U_i. Because a pair of identities of two communication parties are integrated into the one-way hash function to generate the non-interactive session key, and AES encryption algorithm is secure, even the insider legal participants cannot forge a new valid report to impersonate and frame any other innocent residential user. Therefore, the communications from U_i to GW cannot be polluted maliciously. Due to the same reason, the data integrity of the communications from GW to CC can be ensured similarly. In summary, the proposed scheme achieves data integrity throughout the whole communications.
  - Message replay attack resistance

    In the proposed scheme, after receiving the message $< C_i', ID_i >$ from U_i, GW decrypts C_i' to obtain $\mathrm{C_i}||\mathrm{ID_i}||ID_g||g_\tau$. Then, according to the current time point t_τ, GW computes $H_1(t_\tau)$ and checks whether $g_\tau = H_1(t_\tau)$ holds. Because only the fresh package corresponding to the current reporting time point t_τ can pass the verification, the proposed scheme can resist the message replay attack.
- *The user's communication link is secure, reliable, and fault tolerant.*

  We innovate a novel online/offline pre-caching technique to achieve fault tolerance. Even under the circumstances when $\hat{U} \subset U$ do not work, the value r, which is kept secret by CC, together with the pre-cached values of $Y_{t_\tau,\mathrm{i}}$ (for $U_i \in \tilde{U}$), the corresponding secret information of the malfunctioning users $\widetilde{\mathrm{U}}$, still can be used to recover the aggregated measurement of $\sum_{U_i \in U/\tilde{U}} m_i$. Specifically, after AES decrypting and obtaining $\tilde{C}_g$, CC utilizes the secret information $\overline{C}_\tau$= $\prod_{U_i \in \widetilde{\mathrm{U}}} Y_{\mathrm{t_\tau,i}}$ and r to compute $\tilde{C}_g \cdot (\overline{\mathrm{C}}_\tau{}^r) \cdot (h_\tau^{-r}) = g_\tau{}^{\sum_{U_i \in U/\tilde{U}} m_i}$, and recovers the sum usage of functioning smart meters $\sum_{U_i \in U/\tilde{U}} m_i$ ultimately. Without the secret information r and $\overline{\mathrm{C}}_\tau$, anyone else cannot recover the sum of functioning smart meters' usage data, not to mention each user's private usage data m_i.

6 Performance evaluation

Table 2 Comparisons of computation complexity

Protocol	User (ms)	Aggregator (ms)
Ours	4.9706	$0.15n + 3.0874$
[4]	10.2176	$3.8n + 6.7738$
[21]	123.9996	$219.3038n - 181.3$
[22]	19.7844	$0.02n + 38.1518$
[25]	38.1866	$0.0088n + 38.3088$
[19]	5.59404	$5.59404n$
[7]	5.3538	$0.005n^2 + 4.105n + 3.805$

The proposed data aggregation scheme simultaneously

Table 1 Feature comparison

	Ours	He *et al.* [4]	Alsharif *et al.* [21]	Fan *et al.* [22]	Alharbi *et al.* [25]	Fouda *et. al.* [19]	Chen *et. al.* [7]
D:	Yes	Yes	Yes	Yes	Yes	Yes	N
P:	Yes	Yes	Yes	No	Yes	No	Y
F:	Yes	No	No	/	No	No	Y

D: Data integrity
P: Privacy preservation
F: Fault tolerance

achieves privacy preservation, data integrity, and fault tolerance for smart grid communications. In this section, we will mainly compare the performance of our proposed scheme with the state-of-the-art similar symmetric and asymmetric cryptosystems [4, 7, 19, 21, 22, 25]. The major features of these systems are compared in Table 1. It should be noted that the architecture of our scheme is different from all the aforementioned similar works [4, 7, 19, 21, 22, 25] to support fault tolerance. As a result, our comparison is focused on the common parts in computation complexity and communication cost. Besides the above simulation experiments, base on the real data test platform, the correctness and effectiveness of our proposed scheme are also verified in this section.

6.1 Comparison of computation complexity

Table 3 Time cost of operations

Notations	Descriptions	Time Cost
C_a	Addition	≈ 0.004 ms
C_m	Multiplication	≈ 0.16 ms
C_e	Exponentiation	≈ 1.7 ms
C_H	Hash	≈ 0.0037 ms
C_{AES_E}	AES Encryption	≈ 75 MiB/Second
C_{AES_D}	AES Decryption	≈ 75 MiB/Second
C_p	Pairing	≈ 19 ms
$C_{\text{2-DNF}}$	2-DNF Decryption	≈ 1.06 ms
C_{PK_E}	Public Key Encryption	≈ 0.09 ms
C_{PK_D}	Public Key Decryption	≈ 2.28 ms
C_{HM}	HMAC	≈ 0.00724

We perform the experiments with MIRACL [26, 27] library and JPBC (Java Pairing Based Cryptography) library [28] running on a 3.0 GHz processor Pentium IV system to study the operation cost. For clear illustration, we first abbreviate and enumerate all operation notations, and indicate the time cost of all primitive operations in Table 3. Then, the computation complexity of the user side and the aggregator side of the six schemes are computed and compared in Table 2. Finally, we plot the comparisons of computation complexity in Fig. 4.

It can be seen clearly that our scheme achieves much higher computation efficiency in both user side and aggregator side.

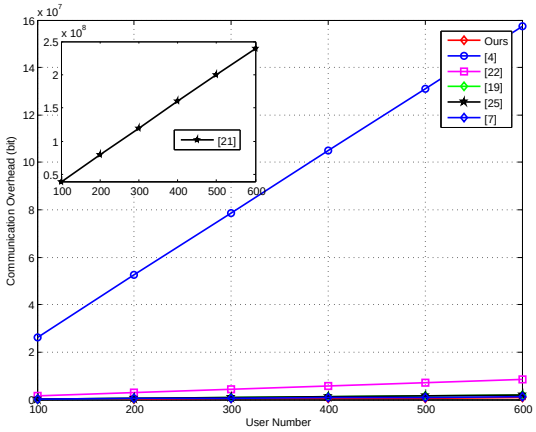


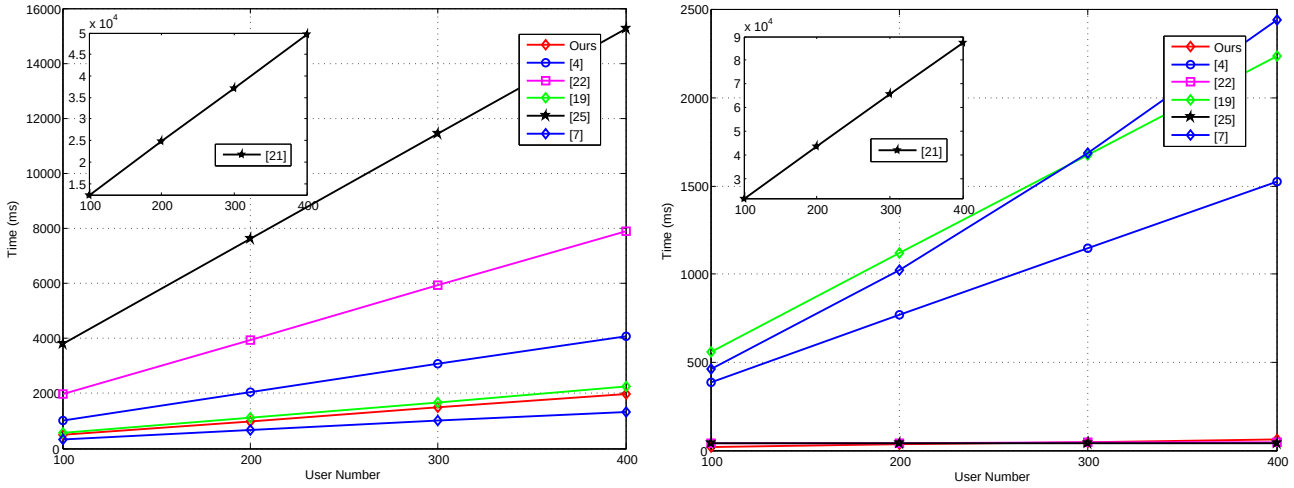
Fig. 5 Performance comparison of communication overhead

6.2 Comparison of communication cost

The communications of the proposed scheme consists of residential users to GW, and GW to CC. By introducing the system-wide trustable entity CC, our proposed scheme achieves enhanced security, which is not considered in other five schemes [4, 7, 19, 21, 22, 25]. Thus, we focus on the common parts of the comparison, *i.e.*, the communication overhead between residential users and GW.

The communication overhead in terms of the number of the users of all the schemes are plotted in Fig. 5. It is obvious that our proposed scheme achieves much lower communication cost compared with the other five schemes.

From the above analysis, our proposed scheme is actually efficient in terms of computation complexity and communi-



(a) Performance comparison of computation cost at user side

(b) Performance comparison of computation cost at aggregator side

Fig. 4 Performance comparison of computation cost

cation cost. Therefore, it is applicable for real-time and high-frequent data aggregation in smart grid communications.

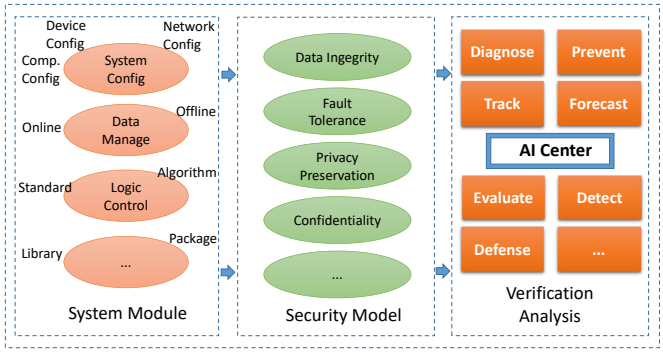


Fig. 6 Algorithm validation based on real data

6.3 Algorithm validation

Besides the above simulation experiments, base on the real data test platform SmartAnalyzer [29], the correctness and effectiveness of the proposed algorithms are verified as shown in Fig. 6. The main verification process and results achieved are as follows.

- *Verification Criteria*
  - The rationality and correctness of the system model, attack model, and design goal set in Section 2 are verified;
  - The security criteria of data integrity, privacy preservation, data damage attack (replay, inject, modify, forge, and/or delay, *etc.*) defence, and malware attack defence are verified;
  - The performance criteria of computation complexity, communication overhead, fault tolerance, and scalability are verified.
- *Verification Processes*
  - The system model considered in Section 2 including four type of entities, *i.e.*, Trusted authority (TA), Control Center (CC), Gateway (GW), and Residential Users (RU), are deployed.
  - The detailed verification of all algorithms in the six phases (System initialization, Data aggregation request, Data aggregation request relay, User data reporting, Secure data aggregation, and Aggregated data recovery) are conducted in the point-to-point fashion.
  - Both the normal cases and abnormal cases (fault conditions and different attacks) are verified.
- *Verification Results*
  - The verification process shows that it achieves all the function, performance, and security requirements defined and set in Section 2.

7 Related works

In this section, we put our emphasis on the discussion of some other literatures [4, 7, 9, 11, 21, 22, 30, 31] related to our research, which also achieve privacy preservation and/or data integrity for smart gird communications. In [7], Chen *et al.* proposed one privacy-preserving data aggregation scheme with fault tolerance for smart grid communications. However, it does not support data integrity. In [9], Chen *et al.* pro-

posed another privacy-preserving multi-functional data aggregation scheme for smart grid communications with multiple data aggregation functions, such as average, variance, and one-way ANOVA. They also extend the basic scheme to resist differential attacks. However, it remains unclear how to support fault tolerance. Li *et al.* proposed one in-network data aggregation architecture for smart grid communications [30]. However, the scheme cannot achieve data integrity. Subsequently, Li *et al.* presented another data aggregation scheme to achieve privacy preservation and data integrity concurrently [31]. The peer-to-peer digital signature exploring homomorphic techniques was designed, and the checksum of the aggregation was calculated and updated along data aggregation flows. However, the hop-by-hop verification process introduced huge additional communication and storage overhead, and the incremental signature verifications launched by the aggregator could expose individual's privacy. In [11], Jongho *et al.* proposed a fault tolerant data aggregation protocol for privacy preserving smart grid communications. The *future ciphertexts* was utilized to support fault tolerance of communication failures, which leads to the heavy round-based communication, computation, and storage overhead. Utilizing the techniques of proxy re-encryption, aggregation signature supporting batch verification, and multi-dimensional matrix computation, Alsharif *et al.* proposed one privacy preserving data collection and access control scheme [21] with multi-recipient for smart grid communications. However, the complex matrix operations in this scheme brought huge communication, computation, and storage overhead, and the scheme does not support fault tolerance. Fan *et al.* put forward one consortium block chain based data aggregation and regulation mechanism for smart grid communications [22] . Based on data signcryption technique, data aggregation with multidimensional information acquisition was achieved. Although this scheme is of some novelty for the introduction of block chain, the performance analysis shows that the length of the signcryption packets turns to be much larger, the description details of the key algorithms of block chain fused secure data encapsulation and communication are too rough, and the fault tolerance attribute is not satisfied either. To resist internal adversaries in smart grid communications, He *et al.* designed a privacy preserving data aggregation scheme [4]. Borrowed the idea of proxy signature, the multi-user registration algorithm was designed in this scheme. In order to report measurements with security properties of data integrity, communication confidentiality, and privacy preservation, the user's private key was generated through the registration procedure. However, after taking a close look at the registration procedure, the user's registration process could be forged, which sows the hazards of impairing data integrity. Although our proposed scheme addresses the similar issues, *i.e.*, to achieve efficient data aggregation with privacy preservation and data integrity in smart grid communications, comparing with existing works, our research emphasis still has some differences: 1) we propose our data aggregation scheme in a more challenging threaten model to resist privacy leakage attack, malware attack, and data damage attack simultaneously; and 2) the enhanced property of fault tolerance is taken into consideration meanwhile. Thus, it additionally improves the reliability and practicability.

8 Conclusions

In this paper, aiming at the practical requirements of smart grid communications, we have proposed a secure data aggregation scheme achieving privacy preservation, data integrity, and fault tolerance simultaneously. The more challenging threaten model is considered, which covers privacy leakage attack, malware attack, and data damage attack. Specifically, utilizing *pre-calculated auxiliary information* techniques, a novel fault tolerant mechanism is designed to aggregate the data of functioning smart meters flexibly and efficiently for any rational number of fault smart meters with arbitrary long failure period. Furthermore, a new efficient authentication technique is proposed to flexibly generate and share session keys in a non-interactive way. The shared session key is utilized for symmetric encryption algorithm to achieve data integrity and source authentication. In addition, through comparative performance analysis, it reveals that the proposed data aggregation scheme outperforms the state-of-the-art similar schemes in terms of computation complexity and communication cost.

References

1. R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.
2. Q. Kong, R. Lu, M. Ma, and H. Bao, "A privacy-preserving sensory data sharing scheme in internet of vehicles," *Future Generation Computer Systems*, vol. 92, pp. 644–655, 2019.
3. Y. Chen, J.-F. Martínez-Ortega, P. Castillejo, and L. López, "A homomorphic-based multiple data aggregation scheme for smart grid," *IEEE Sensors Journal*, vol. 19, no. 10, pp. 3921–3929, 2019.

4. D. He, N. Kumar, S. Zeadally, A. Vinel, and L. T. Yang, "Efficient and privacy-preserving data aggregation scheme for smart grid against internal adversaries," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, 2017.
5. H. Bao and R. Lu, "A lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance," *Peer-to-Peer Networking and Applications*, vol. 10, no. 1, pp. 106–121, 2017.
6. H. Bao, R. Lu, B. Li, and R. Deng, "Blithe: Behavior rule-based insider threat detection for smart grid," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 190–205, 2015.
7. L. Chen, R. Lu, and Z. Cao, "Pdaft: A privacy-preserving data aggregation scheme with fault tolerance for smart grid communications," *Peer-to-peer networking and applications*, vol. 8, no. 6, 2015.
8. H. Bao and R. Lu, "A new differentially private data aggregation with fault tolerance for smart grid communications," *IEEE Internet of Things Journal*, vol. 2, no. 3, pp. 248–258, 2015.
9. 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*, vol. 8, no. 5, pp. 777–792, 2015.
10. E. Shi, H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Annual Network and Distributed System Security Symposium (NDSS)*. Citeseer, 2011.
11. J. Won, C. Y. Ma, D. K. Yau, and N. S. Rao, "Proactive fault-tolerant aggregation protocol for privacy-assured smart metering," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 2804–2812.
12. H. Bao and L. Chen, "A lightweight privacy-preserving scheme with data integrity for smart grid communications," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 4, pp. 1094–1110, 2016.
13. P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.
14. S. H. M. Zargar and M. H. Yaghmaee, "Privacy preserving via group signature in smart grid," in *Proceedings of the Electric Industry Automation Congress (EIAC), Mashhad, Iran*, 2013, pp. 13–14.
15. M. Badra and S. Zeadally, "Design and performance analysis of a virtual ring architecture for smart grid privacy," *IEEE transactions on information forensics and security*, vol. 9, no. 2, pp. 321–329, 2014.
16. W. D. Neumann, "Horse: an extension of an r-time signature scheme with fast signing and verification," in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, vol. 1. IEEE, 2004, pp. 129–134.
17. A. Perrig, "The biba one-time signature and broadcast authentication protocol," in *Proceedings of the 8th ACM conference on Computer and Communications Security*. ACM, 2001, pp. 28–37.
18. D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ecdsa)," *International journal of information security*, vol. 1, no. 1, pp. 36–63, 2001.
19. M. M. Fouda, Z. M. Fadlullah, N. Kato, R. Lu, and X. S. Shen, "A lightweight message authentication scheme for smart grid communications," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, 2011.
20. Y. Ding, B. Wang, Y. Wang, K. Zhang, and H. Wang, "Secure metering data aggregation with batch verification in industrial smart grid," *IEEE Transactions on Industrial Informatics*, 2020.
21. A. Alsharif, M. Nabil, M. M. Mahmoud, and M. Abdallah, "Epda: Efficient and privacy-preserving data collection and access control scheme for multi-recipient ami networks," *IEEE Access*, vol. 7, 2019.
22. M. Fan and X. Zhang, "Consortium blockchain based data aggregation and regulation mechanism for smart grid," *IEEE Access*, vol. 7, 2019.
23. D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *Theory of Cryptography Conference*. Springer, 2005, pp. 325–341.
24. J. Katz, A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*. CRC press, 1996.
25. K. Alharbi and X. Lin, "Lpda: a lightweight privacy-preserving data aggregation scheme for smart grid," in *IEEE International Conference on Wireless Communications and Signal Processing (WCSP)*, 2012.
26. M. Scott, "Miracl-multiprecision integer and rational arithmetic c/c++ library (1988-2007)," *Homepage at http://www.shamus.ie*.
27. R. Li, C. Sturtivant, J. Yu, and X. Cheng, "A novel secure and efficient data aggregation scheme for iot," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1551–1560, 2018.
28. B. Lynn *et al.*, "Pbc: The pairing-based cryptography library," *http://crypto.stanford.edu/pbc*, 2011.
29. M. A. Rahman, E. Al-Shaer, and P. Bera, "A noninvasive threat analyzer for advanced metering infrastructure in smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 273–287, 2012.
30. F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *2010 First IEEE International Conference on Smart Grid Communications*. IEEE, 2010, pp. 327–332.
31. F. Li and B. Luo, "Preserving data integrity for smart grid data aggregation," in *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2012, pp. 366–371.