Peer-to-Peer Netw. Appl. (2017) 10:106–121
DOI 10.1007/s12083-015-0410-7

A lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance

Haiyong Bao[1,2] · Rongxing Lu[1]

Received: 19 May 2015 / Accepted: 24 August 2015 / Published online: 19 September 2015

Abstract To design an efficient and secure data aggregation scheme fitting real applications has been pursued by research communities for a long time. In this paper, we propose a novel secure data aggregation scheme to simultaneously achieve privacy preservation and data integrity with differential privacy and fault tolerance. Specifically, by introducing some auxiliary ciphertext subtly, a novel distributed solution for fault tolerant data aggregation is put forward to be able to aggregate the functioning smart meter measurements flexibly and efficiently for any rational number of malfunctioning smart meters with discretional long failure period. The proposed scheme also achieves a good tradeoff of accuracy and security of differential privacy for arbitrary number of malfunctioning smart meters. In the proposed scheme, a novel efficient authentication mechanism is also proposed to generate and share session keys in a noninteractive way, which is leveraged for AES encryption to achieve source authentication and data integrity of the transmitted data. Furthermore, through decentralizing the computational overhead and the authority of the hub-like entity of the gateway, the security of our proposed scheme is enhanced and the efficiency is improved significantly. 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, communication cost, robustness of fault tolerance, and utility of differential privacy.

Keywords Smart grid · Data aggregation · Privacy preservation · Differential privacy · Fault tolerance

1 Introduction

Comparing with the traditional power grid, smart grid has assimilated various technologies, e.g., data communication and analyzing, sophisticated control and sensing technologies, into the traditional power grid, enables the power distribution to be more efficient and reliable from power generation, transmission, and distribution to end user consumption, and supports the renewable energy [2]. Specifically, as illustrated in Fig. 1, because of smart grid's cyber-physical architecture together with Information and Communications Technology, the control center (CC) can monitor, control, and predict energy generation/consumption in an efficient and real-time way. However, the real-time user data, e.g., collected every 15 min, contain detailed power usage habits which highly correlates to user's privacy, thus they must be protected from unauthorized accesses. In addition to privacy preservation, data integrity is also critical in smart grid communication, otherwise, an attacker could steal or pollute energy usage and consumption information to diminish the availability of smart grid. Therefore, it is of great significance to simultaneously preserve user privacy and assure data integrity in smart grid communications.

This work is an extension of our conference version published in ICC 2015 (London, UK, June 8-12, 2015) [1].

✉ Rongxing Lu
rxlu@ntu.edu.sg

Haiyong Bao
baohaiyong@gmail.com

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

[2] School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

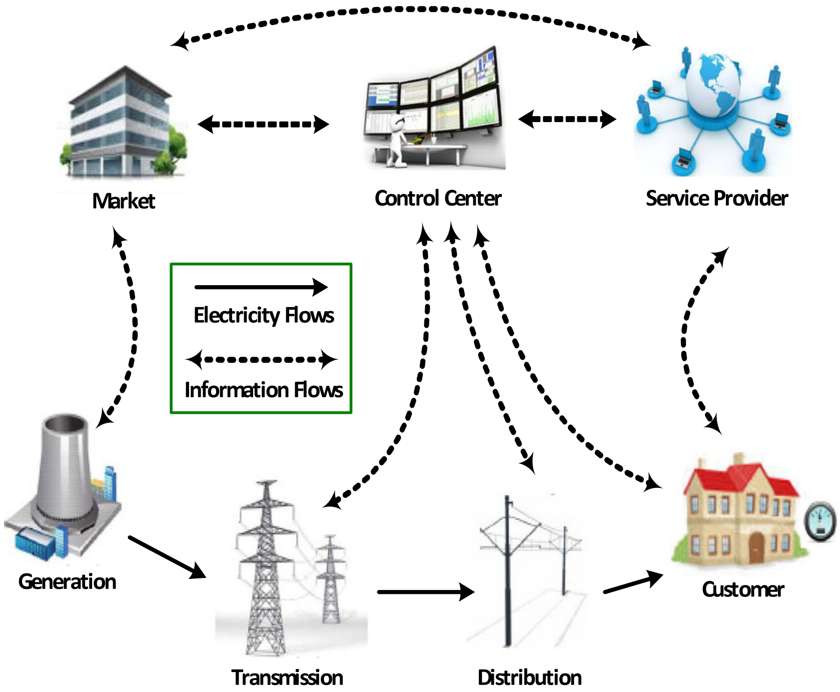


Fig. 1 The conceptual smart grid system architecture

In order to address the privacy issues, several privacy-preserving data aggregation schemes for smart grid communications have been proposed [2–11]. Most of them [2, 3, 6, 7, 10] utilize the homomorphic encryption techniques [12] to encrypt and aggregate users' data in the local area gateway (GW) and forward the data to CC without decryption. However, they only consider protection user's privacy against GW, while CC is still easy to obtain individual user's data. This is because the private key CC holds may not only be used to decrypt the aggregated data, but also be abused to reveal any single user's electricity usage. More seriously, some strong adversaries may deploy undetectable malwares to GW or CC for privacy disclosure of users [3, 10]. This may also conflict users' privacy concerns. When strong adversaries are considered, who aim to snoop user's privacy, these privacy-preserving data aggregation schemes are not robust enough to keep user's privacy unexposed. Other aggregation schemes like [4, 5, 8, 9, 11] take advantage of key management techniques, i.e., the sum of all participants' (including all users and CC) random numbers equals to 0, to diminish CC's authority. One major drawback of such mechanisms is that they are not able to tolerate the report failures [4, 8, 11]. Even though single user fails to report data at some time point, CC would not be able to get anything because the sum of the random numbers in the ultimate encrypted aggregation is no longer 0. This can be a big problem because smart meters, as low-cost devices running in the unprotected environments, are prone to failures. Another challenging problem that secure data aggregation schemes could face comes from differential attack [13]. Specifically, even though one aggregation scheme is secure, once CC obtains the summation of n users and the counterparts of n-1 users, the privacy of the differential one can be inferred to impair user's privacy. This problem has been studied in several works, such as [4, 13–15]. However, most of them do not support fault tolerance, i.e., it is infeasible to be extended to the scenarios where malfunctioning smart meters or communication failures occur. A handful of them [5, 9] support fault tolerance either with unsatisfying utility or with low efficiency, which are not practical, especially in circumstances when the precise number of malfunctioning smart meters are unpredictable and sometimes may be large to some extent.

Meanwhile, in order to prevent malicious adversaries from impairing (e.g., modify, forge, inject, reply and/or delay, etc.) user's usage data report, several message authentication schemes [16–21] have also been presented to ensure data integrity in smart grid communications. Generally, the existing techniques for authenticating communications in smart grid mainly include Hash to Obtain Random Subsets Extension (HORSE) [17], Bins Balls (BiBa)[16], Digital Signature Algorithm (DSA)[18] and MAC/HMAC (hash-based message authentication code)[19]. Comparing with HORSE and BiBa, DSA achieves higher security. However, the computational complexity, especially at the user side is still very heavy [22]. Performance evaluations show that the MAC/HMAC based authentication technique is more efficient than DSA [19]. However, the public key based session key agreement protocol is needed in each round to ensure data integrity, which still poses heavy computational cost and communication overhead for practical applications.

In addition, in bandwidth-intensive and delay-sensitive smart grid communications, especially in user side, efficient mechanism with low computational cost and communication overhead for smart grid communications should be designed to speed up its real applications.

Thus, how to achieve an efficient, secure (with privacy preservation and data integrity simultaneously) data aggregation having enhanced and robust properties (fault tolerant of failures and secure against differentially attack) for smart grid communications still deserves further investigations. In this paper, we propose a novel secure and lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance for smart grid communications. Specifically, the main contributions of this paper are four-fold.

- Firstly, by introducing auxiliary ciphertext subtly, we put forward a novel distributed solution for fault tolerant data aggregation. Unlike most of the existing similar works, which depend on the central trust authority to trace and separate the malfunctioning smart meters from the functioning ones to be able to aggregate the smart meter measurements in case of report failures, our proposed scheme supports fault tolerance of malfunctioning smart meters without the participation and restriction of any external factors. Specifically, utilizing the auxiliary ciphertexts, 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, observing the fact that user's private data may often suffer from differential attacks, our proposed scheme provides differential privacy by adding appropriate noises chosen from Symmetric Geometric distribution to the aggregation data by GW. To the best of our knowledge, most of the existing similar works cannot support differential privacy and fault tolerance at the same time. A handful of literatures trying to address this problem only consider the scenarios that there is small amount (or fixed maximum number) of malfunctioning smart meters to be able to add appropriate noises to support differential privacy. Our scheme supports differential privacy and fault tolerance simultaneously, and achieves a good tradeoff of accuracy and security of differential privacy for arbitrary number of malfunctioning smart meters.
- Thirdly, by integrating a pair of identities and private/public keys of two communication parties, and current time slot for data report, a novel efficient authentication technique is proposed to flexibly generate and share session keys in noninteractive way. The shared session key is leveraged for AES encryption to achieve source authentication and data integrity of transmitted data. The security analysis and performance evaluation indicate that the proposed mechanism can efficiently and effectively prevent the malicious adversary from impairing and polluting (e.g., modify, forge, injection, reply and/or delay, etc.) the transmitted data.
- Finally, through decentralizing the computational overhead and the power of the hub-like entity GW, which is usually with limited computation resources and is semi-trust, the security of our proposed scheme is enhanced and the efficiency is improved significantly. Specifically, only the encryption of the usage data C_i and the auxiliary ciphertext δ_i are aggregated and processed beforehand by at least two users, respectively, can they be reported to GW. In addition, through comparative performance analysis, we demonstrate that our proposed data aggregation scheme outperforms the state-of-the-art similar schemes [4–6] in terms of computation complexity, communication cost, robustness of fault tolerance, and utility of differential privacy.

The remainder of this paper is organized as follows. We first identify the problem formalization which includes system model, attack model and design goal in Section 2, and briefly recall some preliminaries in Section 3. Then, we present our proposed data aggregation scheme in Section 4. Subsequently, the security analysis and performance evaluation are presented in Section 5 and Section 6, respectively. We also discuss related works in Section 7. Finally, we draw our conclusions in Section 8.

2 Problem formalization

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

2.1 System model

Considering real application requirements of smart grid communications, residential users pay great attention to their privacy when reporting measurements to the control center (CC), smart meters as inexpensive home devices, which are often deployed in unprotected environments, may fail to report the usage data, and users' reported data may be tampered with by the malicious adversary due to the unreliable network channel. In this work, we mainly put our emphasis on how to report users' measurements to CC in a secure (with privacy preservation and data integrity simultaneously) and reliable (with fault tolerance) way. Specifically, in our system model, a typical smart grid communication architecture is considered, as shown in Fig. 2, which includes a trusted authority (TA), a control center (CC), a residential gateway (GW), and a huge amount of

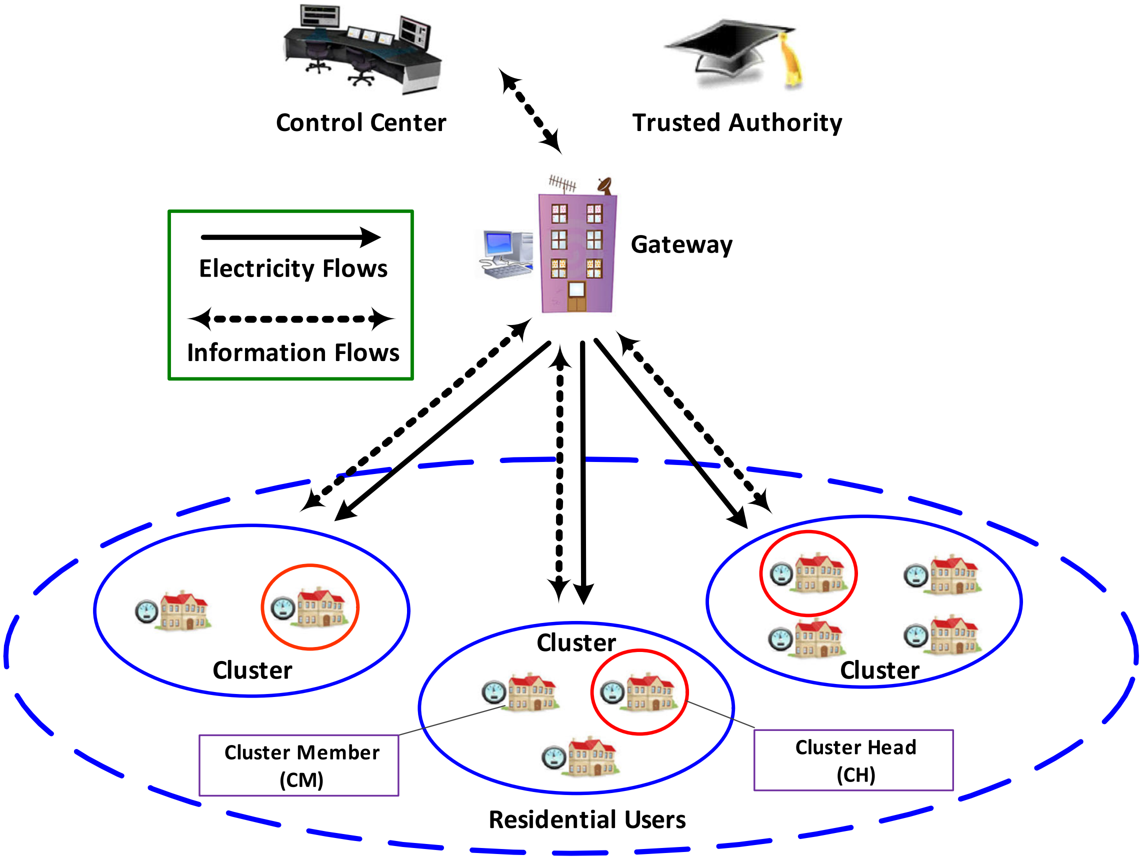


Fig. 2 System model under consideration

residential users $U = \{U_1, U_2, \ldots, U_n\}$ in a residential area (RA).

Trusted authority (TA) TA is a trustable entity who has powerful processing capacity and is in charge of management of the whole system.

Control center (CC) CC is a highly-trusted entity, whose responsibility is to collect, process and analyze the nearly real-time data to be able to provide reliable services for smart grid.

Gateway (GW) GW mainly performs two functions, i.e., aggregation and relaying. The responsibility of aggregation is to aggregate residential users' measurements into a integrated one, and the responsibility of relaying is to help forward the communication flows between CC and residential users in a secure way.

Residential users $U = \{U_1, U_2, \ldots, U_n\}$ Each residential user $U_i \in U$ is equipped with various smart appliances and a smart meter to collect the real-time electricity usage data.

Note that, for the sake of high efficiency and security consideration, our protocol stipulates that any residential user U_i cannot report its measurement individually and directly to GW. Specifically, based on the geographical adjacency, e.g., neighboring districts, U is divided into w clusters $\{\mathrm{CL}_1, \cdots, \mathrm{CL}_w\}$, and the number of users in cluster CL_i is n_i. Here, w is a system parameter, which is dependent on the whole topology and the scale of n. In each cluster $\mathrm{CL}_i \subseteq U$, a cluster head (CH), called U_{h_i}, is appointed. All the other users are cluster members (CMs). Actually, U_{h_i} itself is also a CM of CL_i. There is no particular requirement for appointing the unique CH in each cluster, who can be fixed beforehand or randomly selected temporarily. Then, each CH U_{h_i}, for $i = 1, \cdots, w$, is responsible for pre-aggregating and forwarding all the data reported in CL_i to GW in a certain period, e.g., every 15 min.

2.2 Attack model

In our attack model, CC and GW are considered to be trustable, and the residential users $\{U_1, \cdots, U_n\}$ are "honest" to follow the protocol as well. However, there exists an external adversary $\mathcal{A}$ who may lurk in the RA to eavesdrop the communication flows or intrude into the servers in GW and CC for privacy disclosure of residential users. Besides, $\mathcal{A}$ could also launch some active attacks to impair the data integrity. Specifically, we consider the following most frequently launched attacks intending to divulge user privacy and impair data integrity in smart grid communications:

1) *Privacy divulging attack:* Firstly, an external attacker $\mathcal{A}$ may try to compromise data privacy of a user by eavesdropping the communication package from the user side to the aggregator side. Secondly, $\mathcal{A}$ could maliciously analyze the difference of aggregations between the similar data sets to infer individual ones. Finally, $\mathcal{A}$ may deploy undetectable malwares to GW or CC for privacy disclosure of residential users.
2) *Data alteration attack:* $\mathcal{A}$ may intercept the communication links and impair (modify, forge, inject, reply and/or delay, etc.) users' genuine data contents intended to report.

Besides the above attacks, we also consider that some smart meters could be malfunctioning and in failure status due to meter wastage, physical malfunction, etc., which will fail to report data.

Note that we focus on preventing the external attacks from divulging user's privacy and breaking the data integrity of communication packages. Other attacks, for example, the internal attack, are beyond the scope of this study.

2.3 Design goal

Considering the aforementioned system model and attack model, our design goal is to propose a lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance. Specifically, the following design goals should be achieved:

1) *Privacy preservation:* Firstly, an external attacker $\mathcal{A}$ cannot disclose user's privacy even though $\mathcal{A}$ can eavesdrop the communication flows. Secondly, $\mathcal{A}$ cannot launch differential attack to obtain the individual user's private usage data successfully. Finally, although $\mathcal{A}$ can deploy some undetectable malwares to GW or CC, it still cannot disclose user's private usage data.
2) *Authentication and data integrity:* A user's communication package should be authenticated that it is really transmitted by the corresponding legal residential user. The valid communications cannot be modified during the transmission, i.e., if $\mathcal{A}$ forged, altered, and/or replayed a report, the malicious behaviors should be detected.
3) *Fault tolerance:* The system can still aggregate the data of functioning meters effectively and efficiently even in presence of malfunctioning ones.
4) *Computation efficiency:* The computation efficiency should be achieved in the proposed protocol to support thousands and millions of residential users' data aggregation.

3 Preliminary

In this section, we briefly recall some preliminaries for the construction of our secure and lightweight differentially private data aggregation scheme with fault tolerance.

3.1 Differential privacy

Differential privacy was first proposed by Dwork in [13] in 2006. By adding appropriately chosen noises, e.g., from Symmetric Geometric distribution, Laplace distribution, etc., to the aggregation results, the outputs will become indistinguishable with similar inputs (data sets). We call a randomized algorithm A satisfies ε-differentially privacy, if for any two data sets D_1 and D_2 differing on a single element, for all $S \subset \mathrm{Range}(A)$, $Pr[A(D_1) \in S] \le \exp(\varepsilon) \cdot Pr[A(D_2) \in S]$ holds.

3.2 Differential privacy via symmetric geometric noise

The use of geometric distribution to generate the noise was first put forward by Ghosh et al. in [15]. Specifically, the noise is chosen from the Symmetric Geometric distribution $\mathrm{Geom}(\alpha)$, for $0 < \alpha < 1$, which can be viewed as a discrete approximation of Laplace distribution $Lap(\lambda)$ (where $\alpha \approx \exp(-\frac{1}{\lambda})$). The probability density function (PDF) of the geometric distribution $\mathrm{Geom}(\alpha)$ is $Pr[X = x] = \frac{1-\alpha}{1+\alpha}\alpha^{|x|}$. Formally, if the sensitivity of the aggregation function $A(D)$ is $\Delta A = \max_{D_1,D_2} ||A(D_1) - A(D_2)||_1$ for all data sets D_1 and D_2 differing in at most one element, then by adding geometric noise r randomly chosen from $\mathrm{Geom}(\exp(-\frac{\varepsilon}{\Delta A}))$ to the original aggregation, the perturbed results can achieve ε-differential privacy, i.e., for any integer $k \in Range(A)$, $Pr[A(D_1) + r = k] \le \exp(\varepsilon) \cdot Pr[A(D_2) + r = k]$ holds.

4 Our proposed scheme

In this section, we propose our lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance for smart grid communications. In order to support differential privacy, the encrypted aggregations are perturbed by GW to protect user's privacy against the differential attack. Specifically, because $g_\gamma = H_2(t_\gamma)$ can be always computed by GW for the current time point t_γ, where H_2 is the system-wide public hash function which will be defined in the following *system initialization* Section, after randomly choosing a noise from the geometric distribution, GW could perturb the aggregation by simply multiplying the encrypted noise. In order to achieve ε-differential privacy, for a given ε, according to the sensitivity of aggregation, GW first carefully chooses the parameters of geometric distribution, then computes and adds the noise, which is randomly chosen from the geometric distribution, to the original aggregation to generate the noisy counterparts.

4.1 System initialization

The single trusted authority (TA) is in charge of bootstrapping the whole system in the beginning. Specifically, in the system initialization phase, TA executes the following steps:

- Given the security parameters τ, TA runs the algorithm $\zeta(\tau)$ to obtain the tuple (G, p, h), where G is a cyclic group of prime order p, in which the discrete logarithm problem (DLP) is hard, and $h \in G$ is a random generator of G.
- TA defines two different public cryptographic hash functions $H_1 : \{0, 1\}^* \to G$ and $H_2 : \{0, 1\}^* \to G$.
- TA performs the following steps to assign key materials to residential users $U = \{U_1, U_2, ..., U_n\}$, GW and CC:
    - For each user $U_i \in U$, with the identity ID_i, TA chooses a random number $s_i \in Z_p^*$, assigns it as $U_i's$ private key, and computes $S_i = h^{s_i}$ as $U_i's$ public key.
    - For CC, with the identity ID_c, TA computes $s_0 \in Z_p^*$ satisfying $\sum_{i=0}^{n} s_i = 0 \bmod p$, assigns s_0 as CC′s private key, and computes $S_0 = h^{s_0}$ as CC′s public key.
    - For GW, with the identity ID_g, TA selects a random number $s_g \in Z_p^*$, assigns it as GW′s private key, and computes $S_g = h^{s_g}$ as GW′s public key.
- TA publishes $< G, p, h, H_1, H_2, S_g, S_0, ID_g, ID_c >$ and each $< S_i, ID_i >$, for $U_i \in U$, as the system-wide public information.

Meanwhile, the advanced encryption standard (AES) [23] is adopted in our system as the symmetric cryptosystem. Denote AES-ENC$_k$ and AES-DEC$_k$ as the encryption and decryption algorithms under the symmetric key k, respectively.

4.2 Data aggregation request

Assume the reporting time points of our system, e.g., every 15 min, are defined as $T = \{t_1, t_2, ..., t_{max}\}$ for a sufficient long runtime period. At each time point t_γ, CC launches a *request* for collecting the usage data in the RA as follows:

Step 1: CC computes the hash value $h_\gamma = H_1(t_\gamma)$.
Step 2: CC selects a random $r \in Z_p^*$, and computes $A = h_\gamma^r$.
Step 3: The computed value A is sent to GW by CC.

4.3 Data aggregation request relay

After receiving A from CC, GW relays the data aggregation request by sending $A = h_\gamma^r$ to each $U_i \in U$ in RA, respectively.

4.4 User report generation

Recalling the stipulation in *system model* of Section 2.1 that any residential user $U_i \in U$ cannot report its measurement individually and directly to GW, suppose $U_i \in \text{CL}_i$, and U_{h_i} is the CH of CL_i, then U_i and U_{h_i} perform the following steps collaboratively to report the measurements:

Step 1: Each user $U_i \in \text{CL}_i$ forwards its measurement to U_{h_i} as follows:

- U_i collects its usage data $m_i \in \{0, 1, ..., W\}$ at time point t_γ.
- U_i computes the hash values $g_\gamma = H_2(t_\gamma)$ and $h_\gamma = H_1(t_\gamma)$ for the current reporting time point t_γ.
- U_i encrypts the usage data m_i, using its private key s_i, as $C_i = g_\gamma^{m_i} h_\gamma^{s_i}$ and $\delta_i = A^{s_i} = h_\gamma^{rs_i}$.
- U_i computes the noninteractive session key shared with U_{h_i} as $k_{ih_i} = H_1(S_{h_i}^{s_i} || ID_i || ID_{h_i} || g_\gamma) = H_1(h^{s_{h_i} s_i} || ID_i || ID_{h_i} || g_\gamma)$ and performs AES encryption using k_{ih_i} as $C_i' = \text{AES-ENC}_{k_{ih_i}}(C_i || \delta_i || ID_i || ID_{h_i} || g_\gamma)$.
- U_i sends $< C_i', ID_i >$ to U_{h_i}.

Step 2: After receiving all the messages from each user $U_i \in \text{CL}_i$, U_{h_i} pre-aggregates and reports all users' measurements within CL_i to GW as follows:

- For each received $< C_i', ID_i >$, according to ID_i, U_{h_i} computes the corresponding noninteractive session key shared with U_i as $k_{h_i i} = H_1(S_i^{s_{h_i}} || ID_i || ID_{h_i} || H_2(t_\gamma)) = H_1(h^{s_i s_{h_i}} || ID_i || ID_{h_i} || g_\gamma) = k_{ih_i}$, and decrypts each received C_i' as AES-D EC$_{k_{h_i i}}(C_i') = C_i || \delta_i || ID_i || ID_{h_i} || g_\gamma$.
- Similar as CM $U_i \in \text{CL}_i$, CH U_{h_i} encrypts its own usage data m_{h_i}, as $C_{h_i} = g_\gamma^{m_{h_i}} h_\gamma^{s_{h_i}}$ and $\delta_{h_i} = A^{s_{h_i}} = h_\gamma^{rs_{h_i}}$.
- U_{h_i} pre-aggregates all the encrypted measurements of $U_i \in \text{CL}_i$ including that of itself as $C_{U_i \in \text{CL}_i} =$

$\prod_{U_i \in \mathrm{CL}_i} C_i = g_\gamma^{\sum_{U_i \in \mathrm{CL}_i} m_i} h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i}$ and $\delta_{U_i \in \mathrm{CL}_i} = \prod_{U_i \in \mathrm{CL}_i} \delta_i = \left(h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i}\right)^r$

- U_{h_i} computes the noninteractive session key shared with GW as $k_{h_i g} = H_1(S_g^{s_{h_i}} || ID_{h_i} || ID_g || g_\gamma) = H_1(h^{s_g s_{h_i}} || ID_{h_i} || ID_g || g_\gamma)$, and performs AES encryption using $k_{h_i g}$ as $C'_{U_i \in \mathrm{CL}_i} = \text{AES-ENC}_{k_{h_i g}}(C_{U_i \in \mathrm{CL}_i} || \delta_{U_i \in \mathrm{CL}_i} || ID_{h_i} || ID_g || g_\gamma) = \text{AES-ENC}_{k_{h_i g}}(g_\gamma^{\sum_{U_i \in \mathrm{CL}_i} m_i} h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i} || \left(h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i}\right)^r || ID_{h_i} || ID_g || g_\gamma)$.
- U_{h_i} sets the 1-bit flag F_{h_i} with 1, if all n_i smart meters in CL_i report their measurements correctly. Otherwise, 0 is set to F_{h_i}.
- U_{h_i} reports $< C'_{U_i \in \mathrm{CL}_i}, ID_{h_i}, F_{h_i} >$ to GW.

4.5 Secure report aggregation

Let D be certain subset of residential users, and $A(D) = \sum_{U_i \in D} m_i$, then for any two data sets D_1 and D_2 differing in at most one element, $|A(D_1) - A(D_2)| \le W$ holds, therefore, the sensitivity of A is $\Delta A = W$.

According to the flags F_{h_i}s received from all CHs, GW can judge whether all n smart meters work correctly. If so, after receiving total w number of CH's reports, i.e., $< C'_{U_i \in \mathrm{CL}_i}, ID_{h_i}, F_{h_i} >$, for $i = 1, \cdots, w$, at time point t_γ, GW executes the following steps to aggregate the measurements of all residential users $U_i \in U$:

Step 1: For each received $< C'_{U_i \in \mathrm{CL}_i}, ID_{h_i}, F_{h_i} >$, according to ID_{h_i}, GW computes the corresponding noninteractive session key shared with U_{h_i} as $k_{gh_i} = H_1(S_{h_i}^{s_g} || ID_{h_i} || ID_g || H_2(t_\gamma)) = H_1(h^{s_{h_i} s_g} || ID_{h_i} || ID_g || g_\gamma) = k_{h_i g}$, and decrypts each $C'_{U_i \in \mathrm{CL}_i}$ as $\text{AES-DEC}_{k_{gh_i}} C'_{U_i \in \mathrm{CL}_i} = g_\gamma^{\sum_{U_i \in \mathrm{CL}_i} m_i} h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i} || (h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i})^r || ID_{h_i} || ID_g || g_\gamma = C_{U_i \in \mathrm{CL}_i} || \delta_{U_i \in \mathrm{CL}_i} || ID_{h_i} || ID_g || g_\gamma$.

Step 2: GW aggregates all the received ciphertexts (i.e., multiplying all $C_{U_i \in \mathrm{CL}_i}$, for $i = 1, \cdots, w$), such that the encrypted aggregations of all residential users $U_i \in U$ can be obtained as $C_\gamma = \prod_{i=1}^{w} C_{U_i \in \mathrm{CL}_i} = \prod_{i=1}^{w} g_\gamma^{\sum_{U_i \in \mathrm{CL}_i} m_i} h_\gamma^{\sum_{U_i \in \mathrm{CL}_i} s_i} = \prod_{U_i \in U} C_i = g_\gamma^{\sum_{i=1}^{n} m_i} h_\gamma^{\sum_{i=1}^{n} s_i}$ actually.

Step 3: GW computes the sensitivity of the aggregation as $\Delta A = W$, randomly chooses a noise $\tilde{m}$ from the geometric distribution $\mathrm{Geom}(\exp(-\frac{\varepsilon}{W}))$, and computes the final aggregation as $\tilde{C_\gamma} = C_\gamma \cdot g_\gamma^{\tilde{m}}$.

Step 4: GW computes the noninteractive session key shared with CC as $k_{gc} = H_1(S_0^{s_g} || ID_g || ID_c || g_\gamma) = H_1(h^{s_0 s_g} || ID_g || ID_c || g_\gamma)$, and performs AES encryption using k_{gc} as $\tilde{C_g} = \text{AES-ENC}_{k_{gc}}(\tilde{C_\gamma} || ID_g || ID_c || g_\gamma) = \text{AES-ENC}_{k_{gc}} (g_\gamma^{\sum_{i=1}^{n} m_i + \tilde{m}} h_\gamma^{\sum_{i=1}^{n} s_i} || ID_g || ID_c || g_\gamma)$.

Step 5: GW reports $\tilde{C_g}$ to CC for further computation.

Note that, if smart meters of some users $\hat{U} \subset U$ do not work, i.e., $\hat{U}$ won't report their data at time point t_γ, then GW performs the following steps for privacy-preserving secure report aggregation:

Step 1: For each received $< C'_{U_i \in \mathrm{CL}_i}, ID_{h_i}, F_{h_i} >$, according to ID_{h_i}, GW computes the corresponding noninteractive session key shared with U_{h_i} as $k_{gh_i} = H_1(S_{h_i}^{s_g} || ID_{h_i} || ID_g || H_2(t_\gamma)) = H_1(h^{s_{h_i} s_g} || ID_{h_i} || ID_g || g_\gamma) = k_{h_i g}$, and decrypts each $C'_{U_i \in \mathrm{CL}_i}$ as $\text{AES-DEC}_{k_{gh_i}} C'_{U_i \in \mathrm{CL}_i} = C_{U_i \in \mathrm{CL}_i} || \delta_{U_i \in \mathrm{CL}_i} || ID_{h_i} || ID_g || g_\gamma$.

Step 2: GW aggregates all the received ciphertexts similar as the corresponding procedures when all the smart meters work correctly (i.e., multiplying all $C_{U_i \in \mathrm{CL}_i}$ parts), such that the encrypted aggregations of functioning users $U_i \in U/\hat{U}$ can be obtained as $\hat{C_\gamma} = \prod_{U_i \in U/\hat{U}} C_i = g_\gamma^{\sum_{U_i \in U/\hat{U}} m_i} h_\gamma^{\sum_{U_i \in U/\hat{U}} s_i}$ actually.

Step 3: GW computes the sensitivity of the aggregation as $\Delta A = W$, randomly chooses a noise $\tilde{m}$ from the geometric distribution $\mathrm{Geom}(\exp(-\frac{\varepsilon}{W}))$, and computes the final aggregation as $\tilde{\hat{C_\gamma}} = \hat{C_\gamma} \cdot g_\gamma^{\tilde{m}}$.

Step 4: Similar as *Step 2*, GW also aggregates the additional auxiliary ciphertext of functioning users $U_i \in U/\hat{U}$ as $\delta_\gamma = \prod_{U_i \in U/\hat{U}} \delta_i = (h_\gamma^{\sum_{U_i \in U/\hat{U}} s_i})^r$.

Step 5: GW computes the noninteractive session key shared with CC as $k_{gc} = H_1(S_0^{s_g} || ID_g || ID_c || g_\gamma) = H_1(h^{s_0 s_g} || ID_g || ID_c || g_\gamma)$, and performs AES encryption using k_{gc} as $\tilde{\hat{C_g}} = \text{AES-ENC}_{k_{gc}} (\tilde{\hat{C_\gamma}} || \delta_\gamma || ID_g || ID_c || g_\gamma) = \text{AES-ENC}_{k_{gc}} (g_\gamma^{\sum_{U_i \in U/\hat{U}} m_i + \tilde{m}} h_\gamma^{\sum_{U_i \in U/\hat{U}} s_i} || (h_\gamma^{\sum_{U_i \in U/\hat{U}} s_i})^r || ID_g || ID_c || g_\gamma)$.

Step 6: GW reports $\tilde{\hat{C_g}}$ to CC when malfunctioning smart meters occur.

4.6 Secure report reading

If all n smart meters work correctly, after receiving $\tilde{C_g}$ from GW, at time point t_γ, CC performs the following steps to obtain the aggregation with Geometric noise, which achieves ε-differential privacy:

Step 1: CC computes the noninteractive session key shared with GW as $k_{cg} = H_1(S_g^{s_0} || ID_g || ID_c || H_2(g_\gamma)) = H_1(h^{s_g s_0} || ID_g || ID_c || g_\gamma) = k_{gc}$, and decrypts the received $\tilde{C_g}$ as

$\text{AES-DEC}_{k_{cg}}(\tilde{C}_g) = g_\gamma^{\sum_{i=1}^n m_i+\tilde{m}} h_\gamma^{\sum_{i=1}^n s_i}||ID_g||ID_c||g_\gamma = \tilde{C}_\gamma||ID_g||ID_c||g_\gamma$.

Step 2: CC computes $\tilde{C}_\gamma h_\gamma^{s_0} = g_\gamma^{\sum_{i=1}^n m_i+\tilde{m}} h_\gamma^{\sum_{i=0}^n s_i} = g_\gamma^{\sum_{i=1}^n m_i+\tilde{m}}$.

Step 3: Since $m_i \in \{1, 2, ..., W\}$, we have $\sum_{i=1}^n m_i \leq nW$. By computing the discrete log of $g_\gamma^{\sum_{i=1}^n m_i+\tilde{m}}$, CC can get the noisy aggregation of users' measurements as $M_{sum} = \sum_{i=1}^n m_i + \tilde{m}$ in expected time $O(\sqrt{nW})$ using Pollard's lambda method [24].

Note that, if smart meters of some users $\hat{U} \subset U$ do not work, $\tilde{\hat{C}}_\gamma$ and δ_γ can be obtained at CC by computing k_{cg}, and performing AES decryption as $\text{AES-DEC}_{k_{cg}}(\tilde{\hat{C}}_g) = g_\gamma^{\sum_{U_i\in U/\hat{U}} m_i+\tilde{m}} h_\gamma^{\sum_{U_i\in U/\hat{U}} s_i}||(h_\gamma^{\sum_{U_i\in U/\hat{U}} s_i})^r||ID_g||ID_c||g_\gamma = \tilde{\hat{C}}_\gamma||\delta_\gamma||ID_g||ID_c||g_\gamma$. Then CC performs the following procedures to recover the noisy aggregation of the functioning smart meters:

Step 1: CC computes $\tilde{\hat{C}}_\gamma/\delta_\gamma^{\frac{1}{r}} = g_\gamma^{\sum_{U_i\in U/\hat{U}} m_i+\tilde{m}} h_\gamma^{\sum_{U_i\in U/\hat{U}} s_i}/((h_\gamma^{\sum_{U_i\in U/\hat{U}} s_i})^r)^{\frac{1}{r}} = g_\gamma^{\sum_{U_i\in U/\hat{U}} m_i+\tilde{m}}$.

Step 2: Similar as the corresponding procedures when all smart meters work correctly, the aggregation of the noisy measurements of the functioning smart meters can be recovered as $\sum_{U_i\in U/\hat{U}} m_i + \tilde{m}$ 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 electricity usage privacy is protected from eavesdropping.* Firstly, an adversary $\mathcal{A}$ may reside in RA to eavesdrop the communication flows. Suppose $\mathcal{A}$ has eavesdropped the report from U_i to CH at time point t_γ as $< C_i', ID_i >$. Because user's measurement is encrypted in the AES ciphertext $C_i' = \text{AES-ENC}_{k_{ih_i}}(C_i||\delta_i||ID_i||ID_{h_i}||g_\gamma)$, $\mathcal{A}$ cannot obtain the corresponding plaintext, provided that the session key k_{ih_i} for AES encryption, which is generated cooperatively by utilizing the mutual communication parties' private keys, is secure against $\mathcal{A}$. In the following, we will show that even though the session key k_{ih_i} for some time point, is exposed to $\mathcal{A}$, who still cannot obtain user's specific measurement from AES plaintexts of $C_i = g_\gamma^{m_i} h_\gamma^{s_i}$ and $\delta_i = A^{s_i} = h_\gamma^{rs_i}$. Observing the electricity usage m_i within 15 min 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), the adversary $\mathcal{A}$ is not able to obtain $u_i's$ private key s_i, and cannot recover $U_i's$ usage data m_i. Similarly, the communications from CH to GW, and from GW to CC are of the same form as U_i's report to CH, thus, $\mathcal{A}$ cannot obtain the individual user's usage data via eavesdropping the communication flows. When some smart meters, say $\hat{U} \subset U$, are malfunctioning, because the value of r is kept secret by CC, anyone else cannot use r to recover the sum usage of functioning smart meter as $\sum_{U_i\in U/\hat{U}} m_i + \tilde{m}$, 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 who could obtain the corresponding plaintexts of $\tilde{\hat{C}}_\gamma$ and δ_γ of some time point.
- *The user's electricity usage privacy is protected from malware attack.* Even though the adversary $\mathcal{A}$, after deploying some undetectable malwares into GW or intruding into the database of GW, has stolen the stored data successfully, who could only get the aggregations and ciphertexts of all users' data. Because GW never decrypts any user's electricity usage data, the adversary $\mathcal{A}$ still cannot get any user's private usage data. In addition, $\mathcal{A}$ could also intrude into the database of CC, however, after decryption, the outputs that CC generated are all noisy aggregations of users' data, which do not reveal individual user's usage data at all. Therefore, the individual user's report is protected from malwares attack.
- *The user's electricity usage privacy is protected from differential attack.* For a given privacy level ε, GW perturbs the aggregations without decryption by adding appropriate geometric noises in the form of ciphertext. By this means, ε-differential privacy is achieved. Specifically, GW adds the noise $\tilde{m}$, which is chosen from $\text{Geom}(\exp(-\frac{\varepsilon}{W}))$, to the exact aggregation to obtain the perturbed one. Assume the adversary $\mathcal{A}$ acquires two perturbed aggregations $s + \tilde{m}^{(s)}$ and $t + \tilde{m}^{(t)}$, where s and t are two aggregations of the two data sets differing in at most one element, respectively, while $\tilde{m}^{(s)}$ and $\tilde{m}^{(t)}$ are the two corresponding geometric noises. Similar to the deduction in [10], since $|s - t| \leq W$, for any integer k, we have $\mu = Pr[s + \tilde{m}^{(s)} = k]/Pr[t + \tilde{m}^{(t)} = k] = Pr[\tilde{m}^{(s)} = k - s]/Pr[\tilde{m}^{(t)} = k - t] = (\frac{1-\alpha}{1+\alpha}\alpha^{|k-s|})/(\frac{1-\alpha}{1+\alpha}\alpha^{|k-t|}) = \alpha^{|k-s|-|k-t|}$. Since $-|s - t| \leq |k - s| - |k - t| \leq |s - t|$ and $0 < \alpha < 1$, we have $e^{-\varepsilon} = (e^{-\frac{\varepsilon}{W}})^W \approx \alpha^W \leq \alpha^{|s-t|} \leq \mu \leq \alpha^{-|s-t|} \leq \alpha^{-W} \approx (e^{-\frac{\varepsilon}{W}})^{-W} = e^\varepsilon$. Thus, ε-differential privacy is satisfied. Therefore, even though the

adversary $\mathcal{A}$ obtains the aggregations of two similar data sets, via launching differential attacks, the individual user's privacy is still not leaked at all.

- *The users' electricity usage data aggregation is secure and reliable with the function of fault tolerance of report failures.* We innovate a new distributed method to realize fault tolerance of users' report failures. Even under the circumstances when $\hat{U} \subset U$ do not work, the value r, which is kept secret by CC, together with the aggregated auxiliary ciphertext δ_γ still can be used to recover the aggregated data of $\sum_{U_i \in U/\hat{U}} m_i + \hat{m}$. Specifically, CC, after decrypting and obtaining $\tilde{\tilde{C}}_\gamma$ and δ_γ, uses the private information of r to recover the sum usage of functioning smart meters. Because r is kept secret by CC, without it, anyone else cannot recover the sum of functioning smart meters' usage data, not to mention each user's private usage data m_i.
- *The user's communication link is protected from data alteration attack.* We will show that user's report can be authenticated that it is really sent by a legal residential user, and the communications cannot be altered during the transmissions.

  - Source Authentication

    Firstly, we consider the communications from CM U_i to CH. U_i first generates the noninteractive session key shared with CH as $k_{ih_i} = H_1(S_{h_i}^{s_i}||ID_i||ID_{h_i}||H_2(t_\gamma))$ using U_i's secret key s_i and CH's public key S_{h_i}. Then the report is encrypted using the session key k_{ih_i}, and the generated ciphertext together with U_i's identity ID_i are transmitted to CH. After receiving the report, according to ID_i, CH computes the same noninteractive session key shared with U_i as $k_{h_i i} = H_1(S_i^{s_{h_i}}||ID_i||ID_{h_i}||H_2(t_\gamma)) = k_{ih_i}$, then uses $k_{h_i i}$ to decrypt the report properly. It is obvious that only if the report came from the legal CM, can it be decrypted correctly, thereby, the source authentication can be ensured. Due to the same reason, the source authentication of the report from CH to GW and from GW to CC can be ensured similarly.

    In summary, the proposed scheme achieves source authentication in the whole communications.
  - Data Integrity

  1) Communications pollution attack resistance

     Firstly, we consider the communications from U_i to CH. Upon receiving $< C_i', ID_i >$, according to ID_i, CH first computes the noninteractive session key $k_{h_i i} = H_1(S_i^{s_{h_i}}||ID_i||ID_{h_i}||H_2(t_\gamma)) = H_1(h^{s_i s_{h_i}}||ID_i||ID_{h_i}||g_\gamma) = k_{ih_i}$ shared with U_i. Then, CH performs the AES decryption using $k_{h_i i}$ to obtain U_i's report. Because the secret keys s_i and s_{h_i}, which are kept secret by CM U_i and CH U_{h_i}, respectively, are utilized to compute the shared session key $k_{h_i i}$ (k_{ih_i}) collaboratively, the external adversary $\mathcal{A}$ cannot obtain the agreed secret key, nor can it alter the original data encrypted and reported by U_i. Because a pair of identities of two communication parties are incorporated into the one-way hash function to generate the noninteractive session key and AES encryption scheme is secure, even the insider legal participants cannot forge a new valid report to impersonate and frame the innocent residential user. Therefore, the communications from CM to CH cannot be polluted maliciously. Due to the same reason, the data integrity of the report from CH to GW and from GW to CC can be ensured similarly.

     In summary, the proposed scheme achieves data integrity throughout the whole communications.

  2) Message replay attack resistance

     In the proposed scheme, after receiving the message $< C_i', ID_i >$ from U_i, CH decrypts C_i' to obtain $C_i||\delta_i||ID_i||ID_{h_i}||g_\gamma$, then according to the current time point t_γ, computes $H_2(t_\gamma)$ and checks whether $g_\gamma = H_2(t_\gamma)$ holds for each CM. Because only the fresh report corresponding to the current time point t_γ can pass the verification, the proposed scheme can resist the message replay attack.

6 Performance evaluation

The proposed scheme achieves privacy preservation and data integrity simultaneously for secure data aggregation with differential privacy 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 schemes [6, 11, 19]. The major functions and features of these systems are first compared in Table 1. Both [11] and [19] support data integrity for smart grid communications. Meanwhile, privacy preservation is supported by the schemes of [6] and [11]. Because all users' blinding factors should be utilized cooperatively to obtain the sum electricity usage in [6] and [11], they cannot support fault tolerance of malfunctioning smart meters. Observing that the regular peer-to-peer communication architecture is

Table 1 Feature comparison

	Proposed scheme	Fan et al.'s scheme[11]	Fouda et al.'s scheme [19]	Erkin and Tsudik's scheme [6]
D:	Yes	Yes	Yes	No
P:	Yes	Yes	No	Yes
F:	Yes	No	Partial†	No

D: Data integrity

P: Privacy preservation

F: Supporting data aggregation with fault tolerance

† : Because the simplex and generic peer-to-peer communication architecture is considered, it cannot be regarded as having achieved fault tolerance completely

used in [19], even though theoretically it could be trivially applied in the circumstances when malfunctioning smart meters occur, comparisons will show that our proposed scheme achieves remarkable advantages in terms of efficiency. Because most of the computations of the *hub-like* entity GW are decentralized to CHs in RA, our proposed scheme is very efficient and scalable to support thousands even millions of residential users' data aggregation. Thus, our comparison is focused on data report in user side, which includes both computation cost and communication overhead. In addition, different from all the aforementioned similar works [6, 11, 19], differential privacy and fault tolerance are taken into consideration at the same time in our scheme. As a result, we also compare our proposed scheme with the state-of-the-art data aggregation schemes [5], which supports differential privacy and fault tolerance, in terms of utility of differential privacy and robustness of fault tolerance.

- *Comparison of computation complexity*

  For our proposed scheme, when a CM U_i reports the measurement, it requires 1 multiplication, 3 hash, and 4 exponentiation operations to compute the ciphertext of usage data and non-interactive session key shared with CH, in which 1 exponentiation operation can be pre-computed beforehand in off-line phase. And 1 AES encryption operation is cost to encrypt the report as well. Suppose the size of the cluster CL_i is n_i, after receiving the ciphertexts from n_i-1 CMs, CH first computes all non-interactive session keys shared with each CM within the same cluster CL_i, and decrypts the received ciphertext. It will cost n_i-1 exponentiation, n_i+1 hash, and n_i-1 AES decryption operations, in which n_i-1 exponentiation operations can be pre-computed beforehand in off-line phase. Subsequently, 3 exponentiation, 2n_i-1 multiplication operations are needed to encrypt CH's usage data and aggregate with the other n_i-1 CMs' reports. Finally, 1 AES encryption operation is cost to achieve data integrity. After receiving all the reports from w CHs, GW first computes the noninteractive session key shared with each CH, and decrypts each report. It will cost w+1 hash, w exponentiation, and w AES decryption operations, in which w exponentiation can be pre-computed beforehand in off-line phase. Subsequently, 2(w-1) multiplication operations are cost to aggregate all received residential users' ciphertexts. Finally, 1 AES encryption operation is cost to achieve data integrity. Because the entity CC, who is in charge of monitoring and controlling the whole smart gird, is introduced in our system model, and it is not considered in the other three schemes [6, 11, 19], the corresponding comparisons are not performed here.

  In the scheme of [19], each user reports the data to the aggregator in peer-to-peer way. Firstly, 1 public key encryption, 1 hash, 1 public key decryption, and 2 exponentiation operations are needed for each user to agree on the session key with the aggregator. Then, 1 HMAC generation, and 1 AES encryption operations are cost for each user to transmit the authenticated report to the aggreagator. Subsequently, it will also cost 1 public key encryption, 1 hash, 1 public key decryption, and 2 exponentiation operations for the aggregator to agree on the session key with each user, and cost 1 AES decryption, and 1 HMAC verification operations for the aggregator to verify the integrity of each user's report. For n users, n sets of such operations are needed.

  In the scheme of [11], firstly, for each residential user, it costs 3 hash, 2 exponentiation, 1 addition, and 1 multiplication operations to generate the registration information. Then, 1 hash, 2 multiplication, and 3 exponentiation operations are needed to encrypt the measurement, and 1 exponentiation and 1 hash operations are cost to sign on the measurement. Subsequently, n+1 pairing, 3n-2 multiplications, 2n+2 exponentiation, n+1 hash, and 1 2-DNF formulas cryptosystem decryption operations are needed for the aggregator to aggregate and recover the sum of usage data.

Table 2 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_{HM}	HMAC	≈138 MiB/Second
$C_{\mathrm{HM_V}}$	HMAC verification	≈138 MiB/Second
$C_{\mathrm{AES_E}}$	AES encryption	≈ 75 MiB/Second
$C_{\mathrm{AES_D}}$	AES decryption	≈ 75 MiB/Second
$C_{\mathrm{PK_E}}$	Public key encryption	≈ 0.09 ms
$C_{\mathrm{PK_D}}$	Public key decryption	≈ 2.28 ms
C_p	Pairing	≈ 19 ms
$C_{\text{2-DNF}}$	2-DNF formulas cryptosystem decryption	≈ 1.06 ms

In the scheme of [6], firstly, for each residential user, it costs 1 multiplication, $2n$ addition, and 2 exponentiation operations to exchange random numbers and encrypt the individual measurement, where n is the total number of users. Then, n-1 multiplication and 1 2-DNF formulas cryptosystem decryption operations are needed for the aggregator to aggregate and recover the sum usage data.

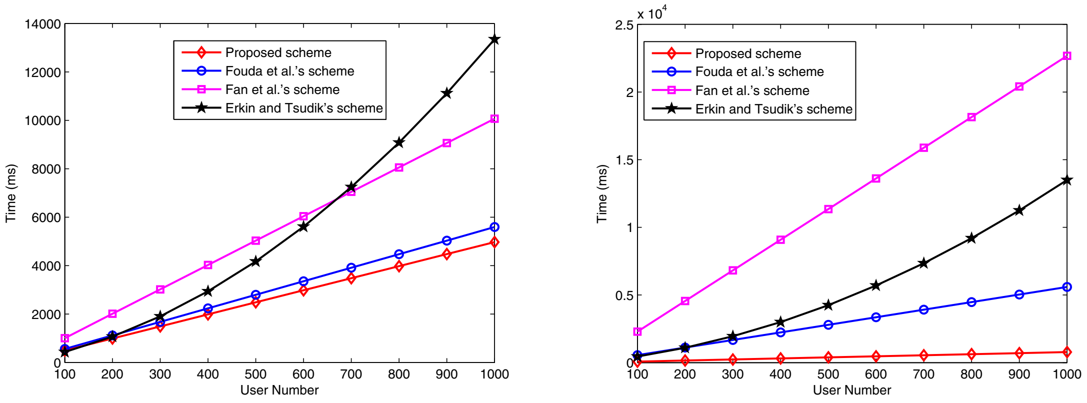
Note that one of the remarkable advantages of our proposed scheme over the other three schemes [6, 11, 19], lies in the characteristic of decentralizing the computational cost of the *hub-like* entity GW. The number of the total clusters w and the scale of each cluster n_i, for $i = 1, \cdots, w$, are configurable. Here, for quantitative comparison, without loss of generality, we set all parameters of n_i to 10 in our proposed scheme, where $i = 1, \cdots, w$, and thus w equals to $\lceil n/10 \rceil$. For clear illustration, we first abbreviate all the operation notations as enumerated in Table 2. Then, the computation cost of each user and the aggreagtor (for our proposed scheme, the corresponding computation cost includes both CH and GW) of the four schemes are compared in Table 3. Furthermore, we perform the experiments with MIRACL [25, 26] library and JPBC (Java Pairing Based Cryptography) library [27, 28] running on a 3.0 GHz processor Pentium IV system to study the operation cost. The time cost of all the primitive operations are indicated in Table 2 as well. Based on the test results, we compare the computation cost of the four schemes in Fig. 3. It can be seen from the figure that, compared with the other three schemes, our proposed scheme greatly reduces the computation complexity in both user side and aggregator side.

- *Comparison of communication cost*

  The communications of the proposed scheme consists of three parts, i.e., CM to CH, CH to GW, and GW to CC. Because via introducing the system-wide trustable entity CC, our proposed scheme achieves enhanced security characteristics which are not considered in other three schemes [6, 11, 19], here, we focus on the comparison of the common parts, i.e., communication overhead between residential users and GW.

Table 3 Computation cost comparisons

Protocol	User	Aggregator
Proposed scheme	$3C_H + C_m + 3Ce + C_{AES_E}$	$nC_{AES_D} + (1.9n - 2)C_m + (1.2n + 1)C_H + 0.3nC_e + (0.1n + 1)C_{AES_E}$
Fouda et al.'s scheme [19]	$2C_e + C_{PK_E} + C_{PK_D} + C_H + C_{HM} + C_{AES_E}$	$n(2C_e + C_{PK_E} + C_{PK_D} + C_H + C_{AES_D} + C_{HM_V})$
Fan et al.'s scheme [11]	$6C_e + C_a + 3C_m + 5C_H$	$(3n - 2)C_m + (n + 1)C_p + (2n + 2)C_e + (n + 1)C_H + C_{\text{2-DNF}}$
Erkin and Tsudik's scheme [6]	$C_m + 2C_e + 2nC_a$	$n(C_m + 2C_e + 2nC_a) + (n - 1)C_m + C_{\text{2-DNF}}$



(a) Performance comparison of computation cost at user side

(b) Performance comparison of computation cost at aggregator side

Fig. 3 Performance comparison of computation cost

In our proposed scheme, when CM reports the measurement to CH, the data report is in the form of $< C_i', ID_i >$, where $C_i' =$ AES-ENC$_{k_{ih_i}}(C_i||\delta_i||ID_i||ID_h||g_\gamma)$ and ID_i is CM's identity. Each CM's individual communication overhead will be 1684 bits, if AES-256 encryption is chosen, and the length of ID_i and ID_{hi}, and the security parameters of τ, are set to 20-bit, 20-bit, and 512-bit, respectively. Each CH U_{h_i} collects the reports from total n_i-1 CMs, which indicates that the total communication overhead between CMs and CH is 1684(n_i-1) bits in one cluster. For all w clusters, the overall communication overhead will be $\sum_{i=1}^{w} 1684(n_i\text{-}1)$ bits totally. Subsequently, each CH aggregates and forwards the report to GW in the form of $< C'_{U_i \in \mathrm{CL}_i}, ID_{h_i}, F_{h_i} >$, where $C'_{U_i \in \mathrm{CL}_i} =$ AES-ENC$_{k_{h_i g}}(C_{U_i \in \mathrm{CL}_i}||\delta_{U_i \in \mathrm{CL}_i}||ID_{h_i}||ID_g||g_\gamma)$, ID_{h_i} is CH's identity, and F_{h_i} is a flag with 1-bit length. Similarly, each CH's individual communication overhead is 1685 bits, and the overall communication overhead between CHs and GW will be 1685w bits for w pairs of communications. In summary, the overall communication overhead of our proposed scheme for all n residential users to report the measurements to GW will be $\sum_{i=1}^{w} 1684(n_i\text{-}1)+1685w = 1684\sum_{i=1}^{w} n_i + w = 1684n + w$ bits.

In the scheme of [19], firstly, 2 public key ciphertexts and 1 q-bit element in Z_q^* are needed to be transmitted between each user and the aggregator to agree on the session key k_i. The communication overhead will be (2048+512)-bit length, if q is set to 512-bit length, and 1024-bit RSA modulus is chosen. Then, each user U_i reports the measurement in the form of ENC$_{k_i}(m_i||T_i||HMAC_{k_i}(m_i||T_i))$ to the aggragator. The size will be 384 bits, if the length of $m_i||T_i$ is set to 200-bit, and AES-256 encryption and HMAC(SHA-1) are chosen. Thus, each user's total communication overhead is 2944 bits. For total n users, the overall communication overhead turns to be 2944n bits.

In the scheme of [11], firstly, $< Y_i, \alpha_i, \beta_i, r_i, ID_i >$ need to be transmitted for each user to register into the system, and the communication overhead will be 2068 bits, if 512-bit q is chosen, and the size of each user's identity is with 20-bit length. Then, each user reports the measurement in the form of $CT_i = g_0^{(H_2(t)h^{r_i'})^{\pi_i}}$ and $\delta_i = H_1(t||CT_i)^{x_i}$ to the aggregator. The size will be 1536 bits, if 1024-bit n is chosen. The aggregator collects the reports from total n users, which indicates that the overall communication overhead between the users and the aggregator is 3604n bits.

In the scheme of [6], users should exchange random numbers and broadcast reports each other in every time point. If 1024-bit modular size is chosen, then the overall communication overhead for all n residential users to repot and aggregate the measurements will be $3n$(n-1) · 1024.

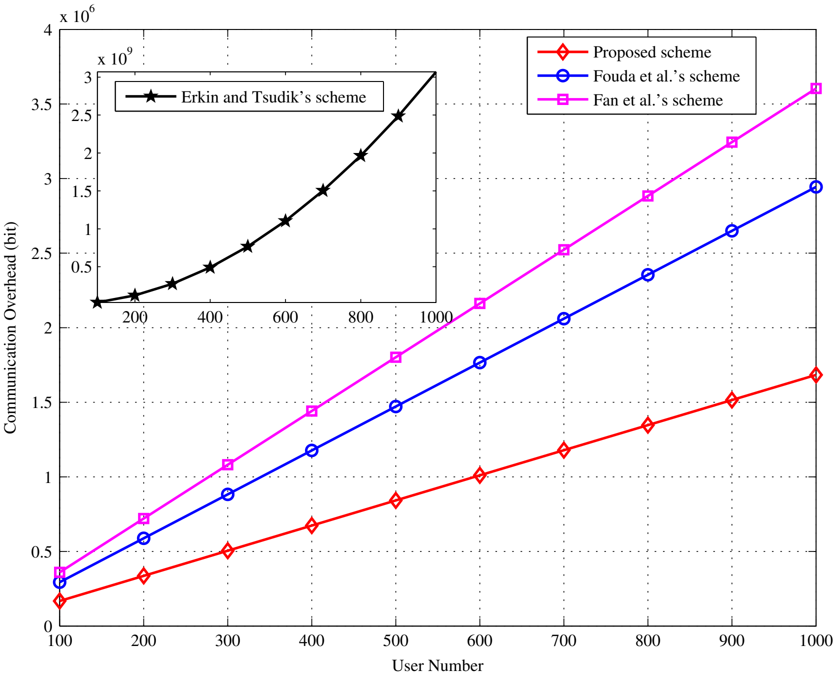
We compare the communication overhead of the four schemes in Fig. 4. It can be seen from the figure that, compared with the other three schemes, our proposed scheme achieves lower communication overhead.

From the above analysis, our proposed scheme is actually efficient in terms of computation complexity and communication cost, which is applicable for the real-time high-frequent data aggregation in smart grid communications.

- *Comparison of robustness of fault tolerance*

Because our scheme inherits all the basic functionality and performance requirements of fault tolerant smart grid data aggregation schemes, here, we focus on the comparison of our scheme with the state-of-the-art data

Fig. 4 Performance comparison of communication overhead

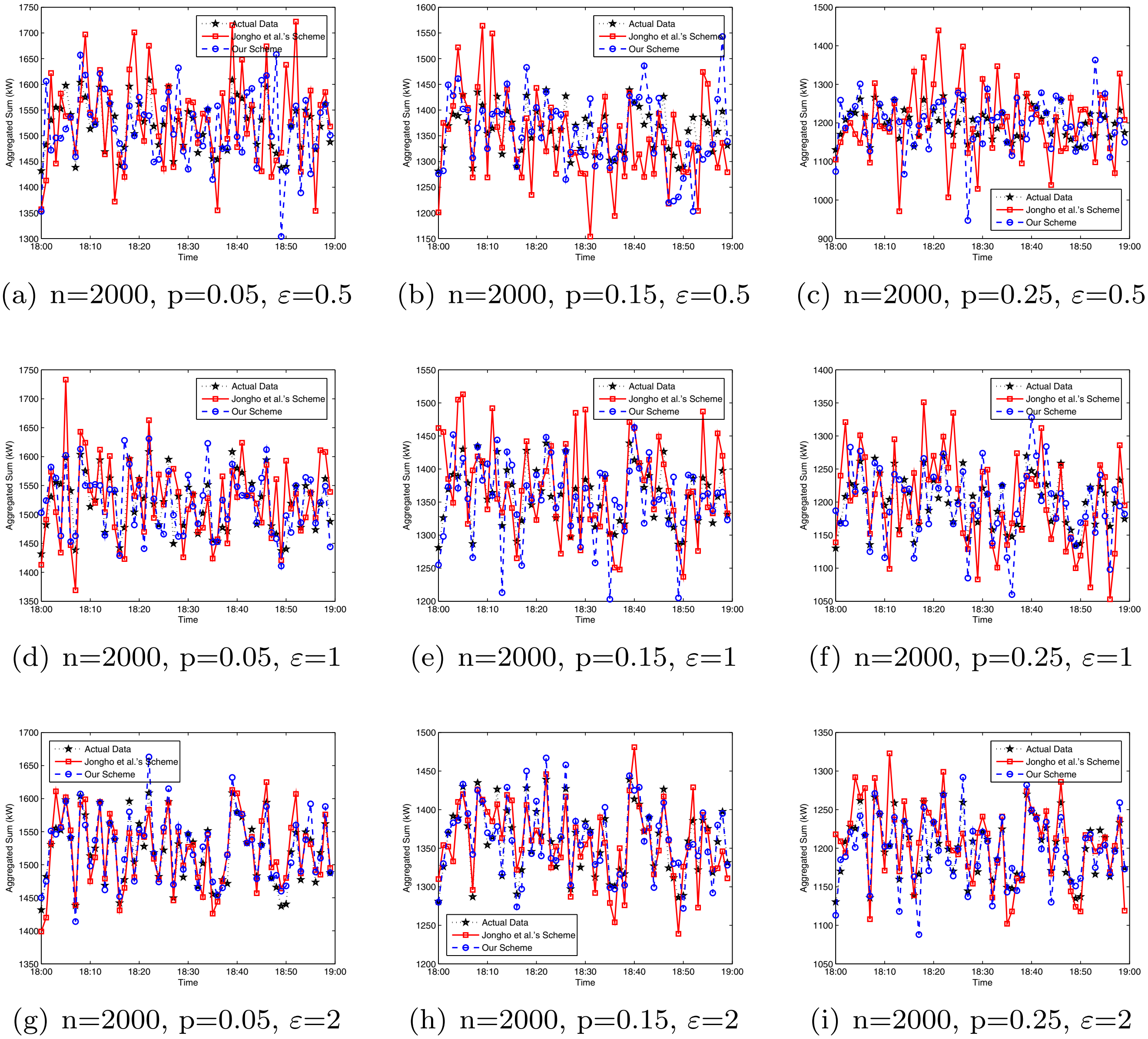


aggregation scheme [5] in terms of robustness of fault tolerance. In the scheme of [5], GW stores B pieces of *future ciphertexts* for each smart meter to support fault tolerance. Without loss of generality, assume the data report interval of the smart grid communications of [5] is T. And suppose at time point T_a, due to some failures, certain smart meter U_i cannot report the measurement successfully to GW, and the fault recovery time point is T_b. Thus, the fault persistence period T_{per} is T_b-T_a. If T_{per} >B·T, until the fault smart meter U_i is to be recovered again on T_b, the system of [5] cannot tolerate the fault any longer after the time point of T_a+B·T, because the pre-stored *future ciphertexts* are used up. The robustness of fault tolerance turns to be much worse when the number of the malfunctioning smart meters increases. In order to support more robust fault tolerance, the system parameter of the buffer size B of [5] should be increased further. However, this causes heavy storage cost, computation complexity and communication overhead. By contrast, our scheme is more robust of fault tolerance and can support efficient data aggregation with any rational number of malfunctioning smart meters with arbitrary long fault period, because our fault tolerance mechanism is not related to the malfunctioning smart meters directly and is independent of any external factors, e.g., *future ciphertexts*.

- *Comparison of utility of differential privacy*

Through the following comparison, we will show that our proposed data aggregation scheme provides higher utility (i.e., low error) in terms of differential privacy than the state-of-the-art data aggregation scheme of [5]. Similar to the scheme of [5], we implement an electricity trace simulator which is able to generate realistic 1-min consumption traces. We produce traces for 2000 households based on this simulator. In the scheme of [5], in order to resit the attack of substracting *current ciphertext* and *future ciphertext*, an additional Laplatics noise $Lap(\lambda)$ is added to each smart meter's *future ciphertext*. However, as commented by the authors of [5], this incurs large errors which increases greatly with the increasing number of malfunctioning smart meters. Our scheme overcomes this drawback, thus, as compared in Fig. 5, it is of better utility than the scheme of [5]. The figures illustrate the traces of the actual total measurements, the noisy counterparts of both [5] and our proposed scheme, for the different parameters, where in each of the figure, n and p denote the total number of the household, and the different ratio of malfunctioning smart meters, respectively. As it can be seen from the figures, the larger the number of p, the more accurate of our scheme comparing with the scheme of [5]. More precisely, let the 1-h root-mean-square-error (RMSE) of Jongho et al.'s protocol [5] and our proposed scheme be γ_1 and γ_2, respectively. The ratios of $\gamma = \frac{\gamma_1}{\gamma_2}$ with p under different privacy level ε are depicted in Fig. 6, which shows that comparing with [5], our proposed scheme always achieves better utility due to much lower errors in each circumstance.

We also set ε, the differential privacy level, to 0.5, 1, 2, for various scenarios with different extent of failures. As it can be seen from Fig. 5, the larger ε is, the smaller noise will be added, and then the utility is higher; while



(a) n=2000, p=0.05, ε=0.5 (b) n=2000, p=0.15, ε=0.5 (c) n=2000, p=0.25, ε=0.5

(d) n=2000, p=0.05, ε=1 (e) n=2000, p=0.15, ε=1 (f) n=2000, p=0.25, ε=1

(g) n=2000, p=0.05, ε=2 (h) n=2000, p=0.15, ε=2 (i) n=2000, p=0.25, ε=2

Fig. 5 Comparison of noisy total consumption between the proposed aggregation protocol and Jongho et al.'s protocol [5]

the smaller ε is, the larger noise will be included, and then the higher level of the privacy can be guaranteed. Compared with the case of $\varepsilon = 2$, the utility in $\varepsilon = 0.5$ is lower, but it is still acceptable. Therefore, in real scenarios, there is a tradeoff between the privacy and utility.

7 Related works

In this section, we put our emphasis on the discussion of some other literatures [5, 6, 11, 19, 20, 29, 30] related to our research which also achieve privacy preservation and/or data integrity for smart gird communications.

Exploring simple cryptographic privacy techniques, Erkin and Tsudik propose one popular privacy-preserving data aggregation scheme without any on-line aggregator or trusted third party [6]. However, users should exchange random numbers and broadcast reports each other in every time point, which incurs both individual and overall communication overhead. Li et al. propose one in-network data aggregation architecture for smart grid communications [29]. Unfortunately, the concrete scheme they designed cannot achieve data integrity. Subsequently, Li et al. present another data aggregation scheme to achieve privacy preservation and data integrity concurrently [30]. The peer-to-peer digital signature exploring homomorphic techniques is designed and the checksum of the aggregation is calculated and

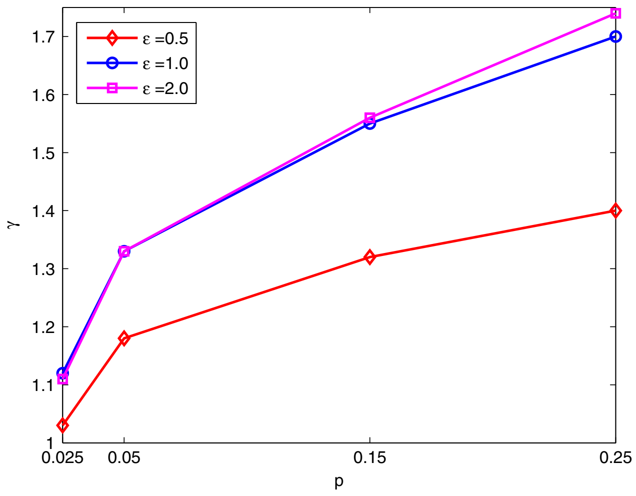


Fig. 6 Comparison of 1-h RMSE between the proposed aggregation protocol and Jongho et al.'s protocol [5]

updated following in-network aggregation flows. However, the hop-by-hop verification process incurs huge additional storage and communication overhead and the incremental signature verifications launched by the aggregator could expose individual's privacy. Fouda et al. propose a message authentication scheme for smart grid communications [19]. Borrowing Diffie-Hellman key exchange technique, the session key is shared between each user and the gateway. Then, with the agreed session key and by employing HMAC technique, the subsequent communications can be authenticated. However, as pointed out by the authors, the security of the system is dependant on the round-based public key encryption and decryption to set up the secure session key, which leads to heavy computation and communication overhead. Besides, the scheme does not achieve privacy preservation for residential user. Alharbi et al. propose one data aggregation for smart grid communications to achieve data security and privacy preservation for residential users [20] with static topology. It is characterized by taking advantage of one-time masking technique to protect user's privacy with high efficiency. However, the session key should be shared between each user and the aggregator, and the neighbouring users also need to agree on session keys, which incurs heavy burden for key management. In [5], Jongho et al. propose a fault tolerant aggregation protocol for privacy-assured smart grid communications. The *future ciphertexts* are leveraged to support fault tolerance of possible communication failures, which leads to the heavy round-based communication, computation and storage overhead. Based on homomorphic encryption techniques, Fan et al. utilize a tree-based aggregation approach to aggregate users' reports efficiently [11]. Through distributing blind factors among all parties including each residential user and the aggregator, the scheme achieves privacy preservation. The registration procedure is interacted between the user and the aggregator to produce user's private key for generating signatures on encrypted data reports to achieve data integrity. However, the pairing based signature verification procedure is resource-consuming. In addition, after taking a close look at the interactive registration procedure, user's private key can be inferred from the public information, 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 divulging attack, which covers eavesdropping attack, differential attack, and malware attack, and data alteration attack simultaneously and 2) we take enhanced properties of differential privacy and fault tolerance into consideration meanwhile; thus, it additionally improves the reliability and practicability.

8 Conclusions

In this paper, we have proposed a secure data aggregation scheme for smart grid communications which not only achieves security properties of privacy preservation and data integrity simultaneously, but also improves the practicability and reliability due to implementation and integration of enhanced properties of differential privacy and fault tolerance. Particularly, under more challenging threaten model which covers communication attack, differential attack, and malware attack, the proposed scheme is secure against privacy divulging attack. Meanwhile, the proposed noninteractive session key agreement mechanism prevents the communications from being polluted and impaired. In addition, through extensive performance evaluation, we have also demonstrated that the proposed scheme outperforms the state-of-the-art similar schemes in terms of computation complexity, communication cost, robustness of fault tolerance, and utility of differential privacy.

Acknowledgments The authors would like to thank the support of Nanyang Technological University under Grant NTU-SUG (M4081196), MOE Tier 1 (M4011177) and AOARD-144029. H. Bao is supported in part by EEE Cybersecurity Research Program, NTU and NSFC program (No. 61100214, No. 61572435).

References

1. Bao H, Lu R (2015) Ddpft: Secure data aggregation scheme with differential privacy and fault tolerance. In: Proceedings ICC 2015. IEEE

2. Lu R, Liang X, Li X, Lin X, Shen X (2012) Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications. IEEE Trans Parallel Distrib Syst 23(9):1621–1631
3. Chen L, Lu R, Cao Z (2014) Pdaft: A privacy-preserving data aggregation scheme with fault tolerance for smart grid communications, Peer-to-Peer Networking and Applications, pp. 1–11
4. Shi E, Chan T-HH, Rieffel EG, Chow R, Song D (2011) Privacy-preserving aggregation of time-series data. NDSS 2(3):4
5. Won J, Ma CY, Yau DK, Rao NS (2014) Proactive fault-tolerant aggregation protocol for privacy-assured smart metering. In: INFOCOM 2014. IEEE, pp 2804–2812
6. Erkin Z, Tsudik G (2012) Private computation of spatial and temporal power consumption with smart meters. Springer, pp 561–577
7. Garcia FD, Jacobs B (2011) Privacy-friendly energy-metering via homomorphic encryption. In: Security and Trust Management. Springer, pp 226–238
8. Rastogi V, Nath S (2010) Differentially private aggregation of distributed time-series with transformation and encryption. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data. ACM, pp 735–746
9. Acs G, Castelluccia C (2011) I have a dream!(differentially private smart metering). In: Information Hiding. Springer, pp 118–132
10. Chen L, Lu R, Cao Z, AlHarbi K, Lin X (2014) Muda: Multifunctional data aggregation in privacy-preserving smart grid communications, Peer-to-Peer Networking and Applications:1–16
11. Fan C-I, Huang S-Y, Lai Y-L (2014) Privacy-enhanced data aggregation scheme against internal attackers in smart grid. IEEE Trans Industrial Informatics 10(1):666–675
12. Paillier P (1999) Public-key cryptosystems based on composite degree residuosity classes. In: Advances in cryptology EUROCRYPT99. Springer, pp 223–238
13. Dwork C (2006) Differential privacy. In: Automata, languages and programming. Springer, pp 1–12
14. Dwork C (2008) Differential privacy: A survey of results. In: Theory and Applications of Models of Computation. Springer, pp 1–19
15. Ghosh A, Roughgarden T, Sundararajan M (2012) Universally utility-maximizing privacy mechanisms. SIAM J Comput 41(6):1673–1693
16. Perrig A (2001) The biba one-time signature and broadcast authentication protocol. In: Proceedings of the 8th ACM conference on Computer and Communications Security. ACM, pp 28–37
17. Neumann WD (2004) Horse: an extension of an r-time signature scheme with fast signing and verification. In: International conference on information technology: Coding and Computing (ITCC 2004), vol 1. IEEE, pp 129–134
18. Johnson D, Menezes A, Vanstone S (2001) The elliptic curve digital signature algorithm (ecdsa). Int J Inf Secur 1(1): 36–63
19. Fouda MM, Fadlullah ZM, Kato N, Lu R, Shen X (2011) A lightweight message authentication scheme for smart grid communications. IEEE Trans Smart Grid 2(4):675–685
20. Alharbi K, Lin X (2012) Lpda: a lightweight privacy-preserving data aggregation scheme for smart grid. In: 2012 International conference on wireless communications and signal processing (WCSP). IEEE, pp 1–6
21. Knox DA, Kunz T (2008) Rf fingerprints for secure authentication in single-hop wsn. In: IEEE international conference on wireless and mobile computing, networking and communications, 2008. WIMOB'08. IEEE, pp 567–573
22. Kgwadi M, Kunz T (2011) Securing rds broadcast messages for smart grid applications. Int J Autonomous and Adaptive Commun Syst 4(4):412–426
23. Daemen J, Rijmen V (2002) The design of Rijndael: AES-the advanced encryption standard. Springer Science & Business Media
24. Menezes AJ, Van Oorschot PC, Vanstone SA (2010) Handbook of applied cryptography. CRC press
25. Scott M (2003) Miracl–multiprecision integer and rational arithmetic c/c++ library, Shamus Software Ltd, Dublin, Ireland
26. Failla P (2010) Privacy-preserving processing of biometric templates by homomorphic encryption, Ph.D. dissertation, Ph. D. dissertation, PhD School in Information Engineering, University of Siena, Italy
27. Scott M (2007) Implementing cryptographic pairings. Lect Notes Comput Sci 4575:177
28. Lynn B, et al. (2011) Pbc: The pairing-based cryptography library, http://crypto.stanford.edu/pbc
29. Li F, Luo B, Liu P (2010) Secure information aggregation for smart grids using homomorphic encryption. In: 2010 First IEEE international conference on Smart Grid Communications (SmartGridComm). IEEE, pp 327–332
30. Li F, Luo B (2012) Preserving data integrity for smart grid data aggregation. In: 2012 IEEE Third international conference on Smart Grid Communications (SmartGridComm). IEEE, pp 366–371

Haiyong Bao, received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai,China, in 2006. He is currently a Post-doctoral Research Fellow with the INFINITUS Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include secure data aggregation, insider attack detection, and applied cryptography.

Rongxing Lu (S09M11SM15) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2012. From May 2012 to April 2013, he was a Post-doctoral Fellow with the University of Waterloo. Since May 2013, he has been an Assistant Professor with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. His research interests include computer network security, mobile and wireless communication security, and applied cryptography. Dr. Lu was the recipient of the Canada Governor General Gold Metal.