Zhang *et al. Journal of Cloud Computing (2023) 12:140*
https://doi.org/10.1186/s13677-023-00522-7

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH **Open Access**

Privacy and integrity-preserving data aggregation scheme for wireless sensor networks digital twins

Zhiming Zhang[1*], Wei Yang[1], Fuying Wu[1] and Ping Li[1]

Abstract

The security technology of digital twin is an important guarantee to ensure the security of digital twin operation, which mainly includes network security technology, data security technology and privacy protection technology. In wireless sensor networks, data aggregation technologies are known as a suitable solution to reduce energy consumption. In addition, due to wireless communications, wireless sensor networks are subject to many attacks. Therefore, it is very important to provide data security in the data aggregation process. In this paper, in order to protect data privacy and verify data integrity, moreover, balance the energy consumption and security during the data aggregation, we present a privacy and integrity–preserving data aggregation scheme for wireless sensor networks based on digital twins technology and homomorphic fingerprinting (HFPIDA). The HFPIDA adopts privacy function to protect data privacy and adopts homomorphic fingerprinting technology to verify the aggregation data integrity. Security analysis shows that the HFPIDA can effectively preserve data privacy and verify data integrity. Simulation results show that the HFPIDA requires less communication and energy overheads, and can achieve higher aggregation accuracy.

Keywords Wireless sensor networks, Data aggregation, Digital twins, Privacy function, Homomorphic fingerprinting

Introduction

Wireless sensor networks (WSNs) are constructed by a large number of sensor nodes in a wireless and multi-hop way. With the rapid development of the Internet of Things, Wireless sensor networks are more widely used in agricultural monitoring, environmental monitoring [1], forest fire detection [2], intelligent transportation, smart home [3], medical monitoring [4], logistics management [5], military and other fields. Because the sensor nodes are limited by calculation, storage and communication, therefore, using data aggregation technology for data transmission can greatly reduce the amount of data transmission in the network, reduce the energy consumption, and extend the life of the whole network.

As most wireless sensor networks are deployed in open environments, they will be attacked by all kinds [6, 7]. Attackers may track, steal or tamper with data forwarded to the base station (BS). Therefore, when data aggregation methods are designed, providing security is very important and challenging [8, 9]. In some applications, the data collected by nodes are sensitive information, so in the process of data aggregation, it is necessary not only to verify the integrity of the data, but also to protect the privacy of the data. Some existing data aggregation methods [10–15] for wireless sensor networks are based on the idea of slicing. The node cuts the data into slices and sends them encrypted, so that the relay node cannot obtain the complete data and realize the protection of data privacy, however, there are more messages exchanges for each node in these methods, which results in high communication overhead. In order to meet the demands of integrity verification and data privacy protection at the same time, some methods [16–18] use a lot

*Correspondence:
Zhiming Zhang
zzm_9650@163.com
[1] School of Software, Jiangxi Normal University, Nanchang 330027, China

of encryption or signature mechanisms with high computational complexity and high communication overhead, which cannot balance the energy consumption and security.

In order to balance the energy consumption and security during the data aggregation, based on homomorphic fingerprinting, a privacy and integrity−preserving data aggregation scheme for wireless sensor networks (HFPIDA) is proposed in this paper. The main contributions of the paper are as follows:

(1) The HFPIDA adopts privacy function and homomorphic fingerprinting technology to protect data privacy and verify the aggregation data integrity. it is mainly to perform hash function operation, fingerprinting function operation and XOR operation. Fingerprinting function operation is essentially a hash function operation, and the computation cost of hash function operation is almost negligible compared with the public key operation used in other schemes, so the HFPIDA is a security and effective scheme, it can balance the energy consumption and security during the data aggregation.

(2) In the HFPIDA, each node only needs to send one packet to its cluster node during the data aggregation. Therefore, compare with the methods based on the idea of slicing, the HFPIDA does not need more messages exchanges, does not generate any redundant data, and it greatly reduces the communication overhead of the network, avoids the data transmission collision and improves the data aggregation accuracy.

The rest of the paper is organized as follows. In Sect. 2, introduces the related work. System model is described in Sect. 3. The HFPIDA scheme is described in Sect. 4. Security analysis is described in Sect. 5. The performance evaluation is implemented in Sect. 6. Section 7 concludes this paper.

Related works

Wireless sensor networks are subjected to many attacks due to wireless communications. Therefore, it is very important to provide data security in the data aggregation process. Scholars have proposed some security and efficient data aggregation schemes.

He et al. [10] proposed a privacy-preserving data aggregation scheme for wireless sensor networks, which included the Slice-Mix-Aggregation privacy protection algorithm (SMART). In the SMART, each node cuts the data into *J* slices and sends (*J-1*) slices to its neighboring nodes, each neighboring node waits for a period time to receive the slices sent by other nodes, then, all the slices perform the mixed calculation and are sent to upper nodes. The SMART preserves the data private with slicing technology, each node cuts its own data and mixes the data slices of neighboring nodes, which increases the difficulty for the attacker to obtain the complete data. However, the SMART has high communication overhead because there are more message exchanges during the data aggregation.

To reduce the communication overhead of the SMART, some improved schemes have been proposed [11–15] based on the idea of slicing. Li et al. [11] proposed a data aggregation privacy protection scheme based on fat tree in wireless sensor networks (FTSMART). For the FTSMART scheme, in the slicing phase, all the nodes need to cut their data into $(n+1)$ slices according to the number n of their parent nodes in the fat tree. In the aggregation phase, each sensor node needs to send one aggregated data packet to the upper node. In the FTSMART, the fat tree is introduced into the data aggregation of wireless sensor network, which has greatly improved the deficiencies of the SMART scheme in the data privacy protection and the aggregation accuracy. Alghamdi et al. [12] proposed a secure data aggregation scheme called sign-share for wireless sensor networks. The network topology is a cluster-based hierarchical structure. Each cluster has two aggregators. Each node divides its data into several slices and sends a part of these data slices to the first aggregator node and another part to the second aggregator. The scheme applies the end-to-end encryption, which can reduce the energy consumption, however, in the data transmission process, if one of the aggregator nodes loses its data for reasons such as attackers, network congestion, and so on, then the data of another aggregator node will be inefficient. Hua et al. [13] proposed an energy-efficient adaptive slice-based secure data aggregation scheme for wireless sensor networks (ASSDA). The network topology is a tree-based structure. In the data slicing process, each sensor node splits data into several slices with different sizes. Then, large-size data slices are transferred to near neighboring nodes and small-size data slices are transmitted to far neighboring nodes, which balances the energy consumption in the network. Zhou et al. [14] proposed an energy-efficient and privacy-preserving data aggregation algorithm for wireless sensor networks (EPDA). The network topology is a tree-based structure. To reduce the communication overhead caused by the data slicing process performed by leaf nodes, an aggregation tree is established between the nodes in the network, and the number of leaf nodes is minimized in the aggregation tree. However, the tree creation process has a high communication overhead. Fang W et al. [15] proposed a novel cluster-based secure data aggregation scheme for

WSNs (CSDA). The network topology is a tree-cluster hierarchical structure. The CSDA uses data slicing technique to protect data privacy, and uses the random pairwise key encryption technique to ensure data security. The CSDA is scalable and improves energy consumption in the network due to applying a tree-cluster hierarchical topology. However, The CSDA has high communication overhead due to using the data slicing technique, and the CSDA uses the hop-by-hop encryption technique, which increases energy consumption.

Parmar P et al. [19] proposed a secure data aggregation protocol using AES in wireless sensor network (SDAPA). The network topology is a tree-based hierarchical structure, each node has two pairwise keys, one shared with its parent node and the other shared with its grandparent node. When a sensor node wants to transmit its data, it sends its data to the parent node and the grandparent node. The grandparent node compares the data received from the child node and the grandchild node, if these values are not the same, the grandparent node rejects the data packets and sends a warning message to the child nodes to retrieve the data correctly. In the SDAPA, the hop-by-hop authentication process is executed. As a result, the malicious node can be quickly removed from the network, but it increases the end-to-end delay and energy consumption in the data transmission process and reduces the network lifetime.

Boubiche D.E et al. [20] proposed a secure data aggregation watermarking-based scheme in homogeneous WSNs (SDAW). The network topology is a cluster-based hierarchical structure, each node sends its data to its cluster head node, and the cluster head nodes aggregate the received data and then forward the aggregated data directly to the base station. The scheme uses a lightweight watermarking technique to secure the network, which can detect fake data packets and isolate malicious nodes. However, it has a high memory overhead due to using a watermarking technique.

Liu X et al. [21] proposed a query privacy preserving for data aggregation in wireless sensor networks (QPPDA). The network topology is a grid-based structure, the whole network is divided into a number of cells. In the QPPDA, the cell member nodes collect the sensed data according to the received query, and encrypt the sensed data using a homophobic encryption technique, then each node sends encrypted data to aggregator node, the aggregator nodes aggregate data received from its cell member nodes and send the aggregated data to the base station. In the QPPDA, the key generation process has high computation overhead by using the homomorphic encryption technique, and it cannot verify the data integrity.

Elhoseny et al. [16] proposed an energy efficient encryption method for secure dynamic wireless sensor networks. The network topology is a cluster-based hierarchical structure, and the clusters are dynamically selected. The scheme uses the elliptic curve cryptography algorithm to generate binary string as encryption keys, and the scheme can prevent the adversary from obtaining the original data. In the meantime, based on the elliptic curve cryptography, Elhoseny et al. [17] proposed a security scheme to protect data privacy for wireless sensor networks. However, the two schemes have high computation overhead due to using the elliptic curve cryptography.

Dener M et al. [18] proposed a secure data aggregation protocol for wireless sensor networks in IoT resistant to DOS attacks. This protocol uses the blowfish encryption algorithm, EAX mode, and RSA algorithm. It can satisfy the often-neglected data availability security requirement and resistant to DOS attacks, however, double encryption/decryption operations occur during data clustering, which increases sensor node's communication load.

Goyal et al. [22] proposed a secure authentication data aggregation scheme for homogeneous underwater wireless sensor networks (SAPDA). The network topology is a cluster-based hierarchical structure. Gateway nodes are tasked to authenticate cluster nodes to ensure that valid cluster nodes manage the clusters. This method has two phases: secure authentication of cluster nodes and secure data aggregation. In this scheme, all sensed data is forwarded to the base station. Hence, it is not scalable because the size of the data packets are increased in each hop. Chenthil T. R. et al. [23] proposed a multi-slot scheduling with a two-layer hexagonal based integrated aggregation approach (MSS-TLHIA) for Underwater Wireless Sensor Networks. In this approach, initially, the entire network is partitioned into several hexagonal grids using the golden ratio. Once the network is partitioned into coverage areas called clusters, a Cluster Head (CH) is selected using the ranking-based fuzzy mechanism. Then, an aggregator node is selected in common for both the layers of the hexagonal grids. Data aggregation is performed using the aggregator node selection process. In order to prevent the energy drain of the aggregator node completely and to prolong their lifetime, the aggregator node is re-selected for every time slot. Furthermore, the occurrence of collision is avoided by the multi-slot scheduling process. The performance of the proposed approach achieves better results in terms of network lifetime, energy consumption and collision rate.

Ozdemir et al. [24] proposed a privacy-preserving data aggregation for wireless sensor networks based on polynomial regression. In this scheme, each node uses the coefficient of polynomial functions instead of the real data, and send the coefficient to the base station, the scheme can protect data privacy and reduce the communication overhead. Based on polynomial regression, Sreenivasulu et al. [25] proposed a non-linear regression model for preserving data privacy in wireless sensor networks.

To sum up, all kinds of current research schemes have their own characteristics. The aggregation schemes [10–15] based on the idea of slicing need to transmit a large number of packets, which will lead to high communication overhead, and these schemes do not take into account the aggregation data integrity. To protect data privacy and verify data integrity, the aggregation schemes [16–18] based on encryption or signature mechanisms need high computational complexity. Hence, it is required to design new data aggregation scheme, which can balance the energy consumption and security.

Preliminaries and system model

Preliminaries

Homomorphic fingerprinting

Hendricks et al. first proposed the homomorphic fingerprinting in [26]. The fingerprinting functions of homomorphic fingerprinting belong to a family of universal hash functions also. Let $IF_{q^{\omega}}$ denote a field of order q^{ω}, Let K be the set of fingerprinting key, and let $P_{q^{\omega}} : K \rightarrow IF_{q^{\omega}}[x]$ be a deterministic algorithm that outputs monic irreducible polynomials of prime degree γ with coefficients in $IF_{q^{\omega}}$, the polynomials are chosen with probabilities taken over the choice of input $r \in K$ uniformly at random, then a fingerprinting function

$$fp(r,d) : K \times IF^{\delta}_{q^{\omega}} \rightarrow IF^{\gamma}_{q^{\omega}} \tag{1}$$

can be defined as

$$fp(r,d(x)) : p(x) \leftarrow P_{q^{\omega}}(r); return(d(x)modp(x)) \tag{2}$$

For any $r \in K$ and $d,d^{'} \in IF^{\gamma}_{q^{\omega}}, b \in IF_{q^{\omega}}$, a fingerprinting function is homomorphic if

$$fp(r,d) + fp\left(r,d^{'}\right) = fp(r,d+d^{'}) \tag{3}$$

and

$$b \cdot fp(r,d) = fp(r,b \cdot d) \tag{4}$$

Let (encode, decode) be a linear erasure code with coefficients $b_{ij} \in IF_{q^{\omega}}$, for $i \in [1,n] and j \in [1,m]$, if $d_1,\ldots,d_n \leftarrow encode^{\delta}(B)$, then for a homomorphic fingerprinting function, the following equation holds

$$fp(r,d_i) = encode^{\gamma}_{i}(fp(r,d_1),\ldots,fp(r,d_m)) \tag{5}$$

where $r \in K$ and $i \in [1,n]$.

Network model

The network model is shown in Fig. 1, the sensor network is composed of sensor nodes, cluster head (CH) nodes and base station. Before deployment, each node j is assigned a random number g_j, a symmetric key $K_{j,BS}$ shared with base station and a public large prime P. After the network is deployed to the target area, all nodes don't move. Adopting the method of reference [27], all nodes are arranged in a cluster-based hierarchical topology. In order to balance the consumption of energy, the cluster head nodes are dynamically selected. Each sensor node sends the collected data to the cluster head node of its cluster. After receiving the sensing data sent by the member sensor nodes, the cluster head node will perform the data aggregation operation, and finally sends the aggregated data to base station. Base station will verify data integrity after receiving all the aggregated data. If the aggregation data is valid, the base station will accept the aggregation data, otherwise it will delete them. As a gateway for external communication, the base station has unlimited computing, storage and communication capabilities, and is absolutely trusted. This paper only considers summation aggregation operation.

Adversary model

It assumes that the sensor nodes and cluster head nodes may be captured except the base station, once a node is captured, the attacker can easily obtain its security information, such as identity, key, etc. Attacker can launch passive attacks or uses the captured malicious nodes to launch active attacks. The specific attacks that attacker can launch are as follows.

(1) By eavesdropping on the communication between nodes, Attacker can obtain the aggregation data sent by the node to the base station, and infer the corresponding original data through these stolen aggregation data, thereby destroying the privacy of the data.
(2) Injecting false data into the network.
(3) Replay attack is launched by stealing packets from nodes.
(4) The captured malicious cluster head node can not only tamper with the aggregation data and destroy the integrity of the data, it can also try to infer the corresponding original data by aggregating the data, thereby destroying the privacy of the data.

This paper does not consider the captured malicious sensor nodes to tamper with their own sensing data, because we think that it is difficult to detect malicious sensor nodes to tamper with their own sensing data only by relying on security protocols, and a small number of

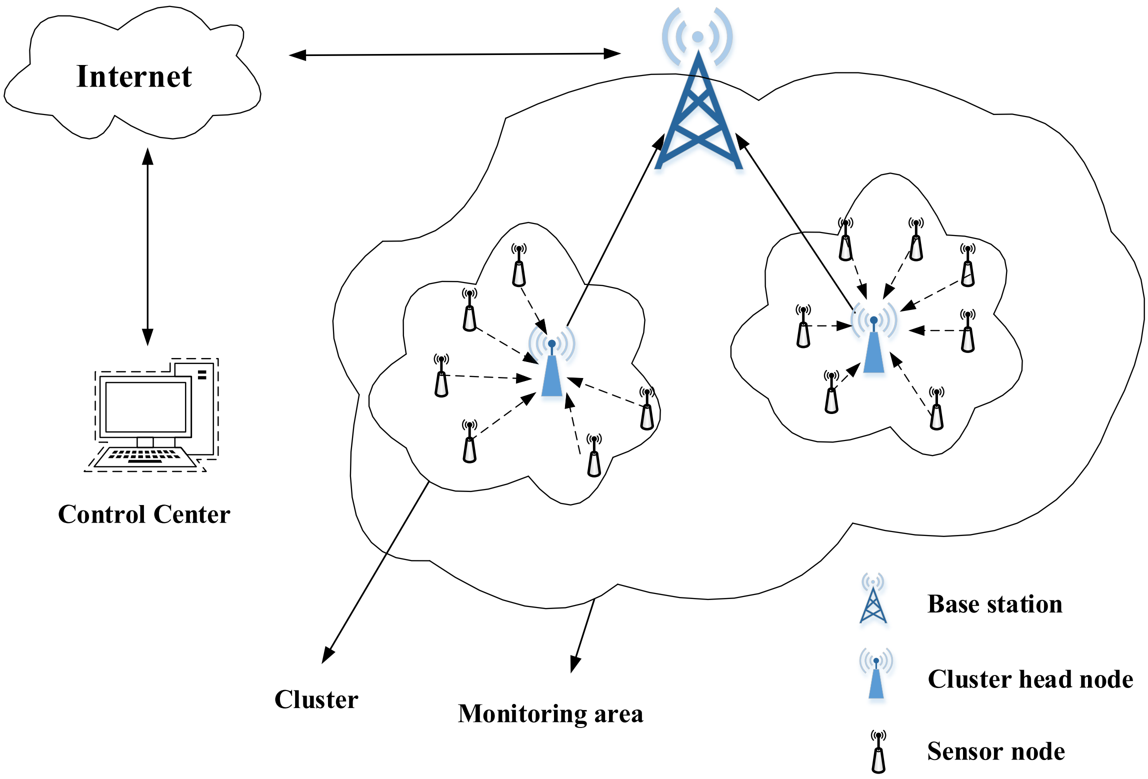

Fig. 1 Network model

captured sensor nodes do not pose a security threat to the network.

Privacy and integrity–preserving data aggregation scheme based on homomorphic fingerprinting: HFPIDA

Every time base station wants to collect the sensing data in the network, it first selects a random number $r, r \in IF_{q^\omega}$ and broadcasts it to all nodes in the network. After a time period, when all nodes receive the random number r, they send the sensing data to the their cluster head node. The cluster head node aggregates the data and sends it to base station. Finally, base station will verify the integrity of the aggregation data. In order to protect the privacy and integrity of data, HFPIDA consists of four steps: privacy data generation, data aggregation, data recovery and verification. This section will describe each step in detail.

Privacy data generation

Suppose that a sensor node j in cluster i senses the data d_j, it first hides the data d_j in a privacy function $f_j(x_j)$, then calculates the homomorphic fingerprinting fp_j of the data d_j as the authentication information of the data, and finally sends the relevant data to the cluster head node CH_i. The specific execution process is as follows.

(1) Sensor node j gets the hash value $h(k_{j,BS})$ of the symmetric key $k_{j,BS}$ shared with base station by the secure one-way hash function $h(.)$, then, sensor node j constructs the privacy function $f_j(x_j)$ with the $h(k_{j,BS})$, rand number g_j, data d_j and a public large prime P as follows.

$$f_j(x_j) = (x_j - h(k_{j,BS}) \oplus g_j) + d_j(mod)P \tag{6}$$

Where $\oplus$ denotes the XOR operation, *mod* denotes modulo operation.

(2) Then it calculates the homomorphic fingerprinting fp_j of the data d_j according to the formula 2 as follows.

$$fp_j = fp(r, d_j) : p(x) \leftarrow P_{q^\omega}(r); return(d(x) mod\ p(x)) \tag{7}$$

The homomorphic fingerprinting fp_j will be used as the authentication information of the data d_j.

(3) Finally, it sends the data $(f_j(x_j), fp_j, g_j)$ to the cluster head node CH_i, where $f_j(x_j)$ denotes the privacy function, fp_j denotes the homomorphic fingerprinting and g_j denotes a rand number.

Data aggregation

Suppose there are m sensor nodes in cluster i. If the cluster head node CH_i receives the data $\{(f_j(x_j), fp_j, g_j), j = 1 \ldots m\}$ sent by all nodes in the cluster, it first aggregates the privacy functions of m sensor nodes in the cluster to obtain the aggregation privacy function $F_i(x_1, x_2, \ldots, x_m)$, then, it aggregates the data authentication information of m sensor nodes to obtain the aggregation homomorphic fingerprinting FP_i, and finally sends the relevant data to the base station. The specific execution process is as follows.

(1) The CH_i aggregates the privacy functions $f_j(x_j)$ of m sensor nodes and gets the aggregation privacy function $F_i(x_1, x_2, \ldots, x_m)$ as follow.

$$F_i(x_1, x_2, \ldots, x_m) = \sum_{j=1}^{m} f_j(x_j) \tag{8}$$

$$= \sum_{j=1}^{m} \left(x_j - h(k_{j,BS}) \oplus g_j\right) + \sum_{j=1}^{m} d_j \ (mod)\ P \tag{9}$$

(2) Then it aggregates the data authentication information fp_j of m sensor nodes and gets the aggregation homomorphic fingerprinting FP_i according to the formula 3 as follow.

$$FP_i = \sum_{j=1}^{m} fp_j \tag{10}$$

$$= \sum_{j=1}^{m} fp(r, d_j) = fp\left(r, \sum_{j=1}^{m} d_j\right) \tag{11}$$

(3) Then it sets $G_i = \{null\}$, and performs set union operation for random number g_j of m sensor nodes to get random number set G_i as follow.

$$G_i = G_i \cup g_j, j = 1 \ldots m, \text{ where} \cup \text{ denotes set union operation} \tag{12}$$

(4) Finally, CH_i sends the data $(F_i(x_1, x_2, \ldots, x_m), FP_i, G_i)$ to the base station, where $F_i(x_1, x_2, \ldots, x_m)$ denotes the aggregation privacy function, FP_i denotes the aggregation homomorphic fingerprinting and G_i denotes random number set in cluster i.

Data recovery

Suppose the whole network is divided into n clusters, each cluster has m sensor nodes. When the base station receives the aggregation data $\{(F_i(x_1, x_2, \ldots, x_m), FP_i, G_i), i = 1 \ldots n\}$ sent by all n cluster head nodes, it performs the following operations to recover the original data.

(1) The base station aggregates the privacy functions $F_i(x_1, x_2, \ldots, x_m)$ of n cluster head nodes and gets the aggregation privacy function $F_{BS}(x_{11}, \ldots, x_{nm})$ as following.

$$F_{BS}(x_{11}, \ldots, x_{nm}) = \sum_{i=1}^{n} F_i(x_1, x_2, \ldots, x_m) \tag{13}$$

$$= \sum_{i=1}^{n} \sum_{j-1}^{m} f_{ij}(x_{ij}) \tag{14}$$

$$= \sum_{j=1}^{n} (\sum_{j-1}^{m} \left(x_{ij} - h(k_{ij,BS}) \oplus g_{ij}\right) + \sum_{j=1}^{m} d_{ij} (mod) P\ ) \tag{15}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} - h(k_{ij,BS}) \oplus g_{ij}\right) + \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} \ (mod)\ P \tag{16}$$

(2) It sets $G_{BS} = \{null\}$, and calculates $G_{BS} = G_{BS} \cup G_i = \{g_{11}, \ldots, g_{nm}\}, i = 1 \ldots n$, where $\cup$ denotes set union operation. Then it takes out each random number g_{ij} from G_{BS} in turn, and finds out the key $k_{ij,BS}$ shared by the corresponding node and base station.

(3) It calculates the independent variable $x_{ij} = h(k_{ij,BS}) \oplus g_{ij}$ of the privacy function $F_{BS}(x_{11}, \ldots, x_{nm})$ according to g_{ij} and $k_{ij,BS}$ in turn. Then it substitutes x_{ij} into the function $F_{BS}(x_{11}, \ldots, x_{nm})$ to recover the original data D_{BS} as following.

$$D_{BS} = F_{BS}(x_{11}, \ldots, x_{nm}) \tag{17}$$

$$= F_{BS}\left(h(k_{11,BS}) \oplus g_{11}, \ldots, h(k_{nm,BS}) \oplus g_{nm}\right) \tag{18}$$

$$= \sum_{i=1}^{n} \sum_{j-1}^{m} \left(h(k_{ij,BS}) \oplus g_{ij} - h(k_{ij,BS}) \oplus g_{ij}\right) + \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} \ (mod)\ P \tag{19}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}d_{ij} \tag{20}$$

Data verification

After recovering the original data D_{BS}, the base station first aggregates the homomorphic fingerprinting FP_i sent by n cluster head nodes to obtain the aggregation homomorphic fingerprinting FP_{BS}, then it calculates the homomorphic fingerprinting $FP\prime_{BS}$ of the recovered original data D_{BS}, and finally verifies data integrity by comparing the results of FP_{BS} and $FP\prime_{BS}$. The specific integrity verification process is as follows.

(1) The base station aggregates the homomorphic fingerprinting FP_i of n cluster head nodes and gets the aggregation homomorphic fingerprinting FP_{BS} according to the formula 3 as following.

$$FP_{BS} = \sum_{i=1}^{n}FP_i \tag{21}$$

$$= \sum_{i-1}^{n}\sum_{j=1}^{m}fp_{ij} = fp(r,\ \sum_{i=1}^{n}\sum_{j=1}^{m}d_{ij}) \tag{22}$$

(2) The base station gets the homomorphic fingerprinting $FP\prime_{BS}$ of the recovered original data D_{BS} by calculating.

$$FP\prime_{BS} = fp(r, D_{BS}) \tag{23}$$

$$= fp(r,\ \sum_{i=1}^{n}\sum_{j=1}^{m}\ d_{ij}) \tag{24}$$

(3) The base station verifies data integrity by comparing the results of FP_{BS} and $FP\prime_{BS}$, if FP_{BS} is equals to the $FP\prime_{BS}$, it accepts the data D_{BS}, otherwise, it means that the data has been tampered with and will not be accepted.

Security analysis

In Sect. 3.3, it introductions that attackers can launch passive attacks or use captured malicious nodes to launch active attacks, which will destroy the privacy and integrity of data. This section will discuss how the HFPIDA scheme proposed in this paper protects the privacy and integrity of data and resists replay attack.

Data privacy analysis

In the HFPIDA, sensor node j hides its data in a privacy function $f_j(x_j)$, that is, the data is encrypted by disturbing the data, and then sent to the cluster head node. Because any intermediate node or attacker has no the key shared by node j and base station, they cannot obtain the sensing data sent by the sensor node to the cluster head node by eavesdropping on the communication between nodes. When the cluster head node CH_i receives the data sent by m nodes in the cluster, it first calculates the aggregation privacy function $F_i(x_1, x_2, \ldots, x_m) = \sum_{j=1}^{m}\left(x_j - h\left(k_{j,BS}\right) \oplus g_j\right) + \sum_{j=1}^{m}d_j(mod)P$, and then sends it to the base station. Any intermediate node or attacker has no the keys shared by m nodes and base station, they cannot obtain the aggregation data $\sum_{j=1}^{m}d_j$ sent by the cluster head node CH_i to the base station by eavesdropping on the communication between nodes. Therefore, the HFPIDA can resist various passive attacks launched by attackers and protect the privacy of single data and aggregation data.

Attackers can capture some sensor nodes or cluster head nodes, so the attackers can obtain the keys and random numbers shared by these captured malicious nodes and base station, and then try to infer the aggregation data $\sum_{j=1}^{m}d_j$ in the aggregation privacy function $F_i(x_1, x_2, \ldots, x_m) = \sum_{j=1}^{m}\left(x_j - h\left(k_{j,BS}\right) \oplus g_j\right) + \sum_{j=1}^{m}d_j(mod)P$ through these keys and random numbers. However, since these captured malicious nodes do not have the keys shared by other sensor nodes and base station, the aggregation data in the privacy function cannot be inferred. Therefore, the HFPIDA can resist the active attacks launched by attackers and protect the privacy of aggregation data.

Data integrity analysis

In the HFPIDA, sensor node j calculates the homomorphic fingerprinting $fp_j = fp\left(r, d_j\right)$ of the data d_j as the authentication information of the data, and cluster head node CH_i calculates the aggregation homomorphic fingerprinting $FP_i = \sum_{j=1}^{m}fp_j$ as authentication information of the aggregation data. The captured cluster head node may tamper with the aggregation data or inject false data, but the base station can find such tampering or injecting false data in the data verification step in Sect. 4.4. Therefore, the HFPIDA can protect the integrity of data.

Resisting replay attack analysis

In the HFPIDA, every time the base station wants to collect the sensing data in the network, it will send a random number $r, r \in IF_{q^{\omega}}$ to all nodes. If the attacker attempts to launch a replay attack by sending the previous data, because the random number r is different every time, and the random r used to calculate the homomorphic fingerprinting $fp_j = fp\left(r, d_j\right)$ is also different, and the base station can find this attack in the data verification step in Sect. 4.4. Therefore, the HFPIDA can resist the replay attack launched by sending the previous data.

The performance evaluation

In this paper, the performances of HFPIDA, SMART and FTSMART are evaluated from the aspects of the communication overhead, the energy consumption and the aggregation accuracy. The simulation experiment environment is carried out on OMNeT + + platform, with 200 nodes randomly distributed in a square area of 400 m × 400 m, the nodes will not move after deployment, and the base station is deployed in the center of the area. The packet size is 128bytes, and the cluster sizes range from 5 to 12. The parameter settings of the experimental simulation are shown in Table 1.

Communication overhead

We adopt the total amount of the packets transmission during data aggregation as a measure of communication overhead.

In the SMART, if each node has M-1 neighboring nodes, each node cuts its data into M slices and sends (M-1) slices to its neighboring nodes in the slicing phase, after mixing, each node sends the new packet to its upper node in the aggregation phase. Therefore, the communication overhead of the SMART is given by

$$CO_{SMART} = N * (M - 1) + N * 1 = N * M \tag{25}$$

Where CO_{SMART} denotes the communication overhead of the SMART, and N denotes the total number of the nodes in the network.

In the FTSMART, the number of each node's parent is different, if a node has n parent nodes, the node cuts its data into (n + 1) slices and sends the (n) slices to its parent nodes in the slicing phase, each node needs to send one packet to its upper node in the aggregation phase. Therefore, the communication overhead of the FTSMART is given by

$$CO_{FTSMART} = \sum_{i=1}^{N} T_i (T_i \epsilon [1, 2, ..., n_{max} + 1]) \tag{26}$$

Where $CO_{FTSMART}$ denotes the communication overhead of the FTSMART, N denotes the total number of the nodes in the network, T_i denotes the amount of the packets generated by node i, and n_{max} denotes the maximum number of the parents for all nodes.

In the HFPIDA, each node only needs to send one packet to its cluster node during the data aggregation. Therefore, the communication overhead of the HFPIDA is given by

$$CO_{HFPIDA} = N \tag{27}$$

Where CO_{HFPIDA} denotes the communication overhead of the HFPIDA, N denotes the total number of the nodes in the network.

We set the number of the neighboring nodes for each node in the SMART is 2, the simulation experiment results of the communication overhead of HFPIDA, SMART and FTSMART are demonstrated in Fig. 2. It can be observed form the Fig. 2 that the communication overhead of SMART is lower than that of SMART and FTSMART. Since each node of HFPIDA does not need to send slices to other neighbor nodes and only send one packet to its cluster, it vastly reduce the communication overhead in the whole network.

Table 1 Simulation parameters

Parameter	Value
Network deployment area (m)	400 × 400
Number of nodes in the network	200
Transmission range(m)	30
Initial energy of each node (J)	4
Transmit power (mw)	50
Receiving power(mw)	10
The simulation time (S)	100

Energy consumption

The amount of energy consumption directly affects the life of the network, so one of the important metrics to demonstrate the performance of the data aggregation scheme is the energy consumption. The energy costs are composed of the cost of transmission, reception and computation. The total energy consumption in an arbitrary node is given by

$$E_{Total} = E_t + E_r + E_c \tag{28}$$

Where E_{Total} denotes the total energy consumption, E_t denotes the energy consumption of transmission packet, E_r denotes the energy consumption of receiving packet and E_c denotes the energy consumption of the computation.

The computation cost of the SMART and FTSMART is mainly to perform slicing operation, encryption and decryption operation, the computational cost of HFPIDA is mainly to perform hash function operation, fingerprint function operation, XOR operation and privacy function addition operation. Fingerprint function operation is essentially a hash function operation, and the computation cost of hash function operation is almost negligible compared with the public key operation used in other schemes, XOR operation is the most basic operations in cryptography. So

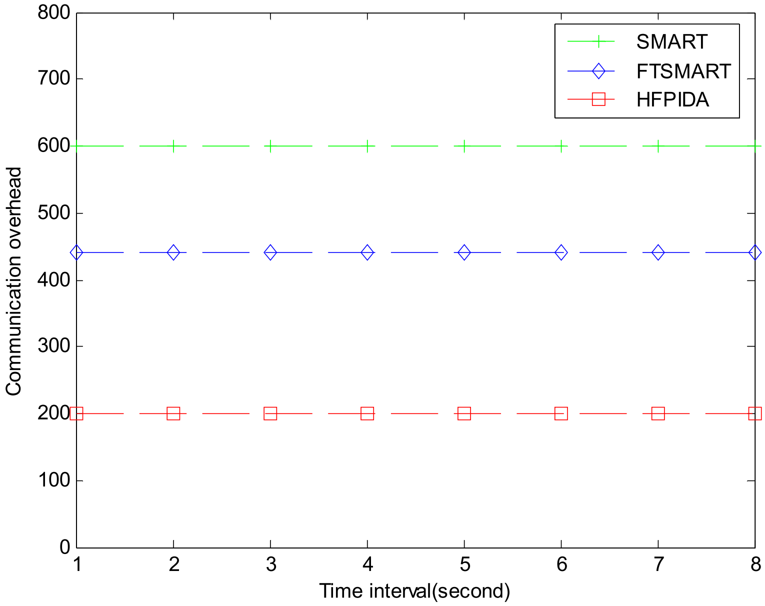

Fig. 2 Communication overhead

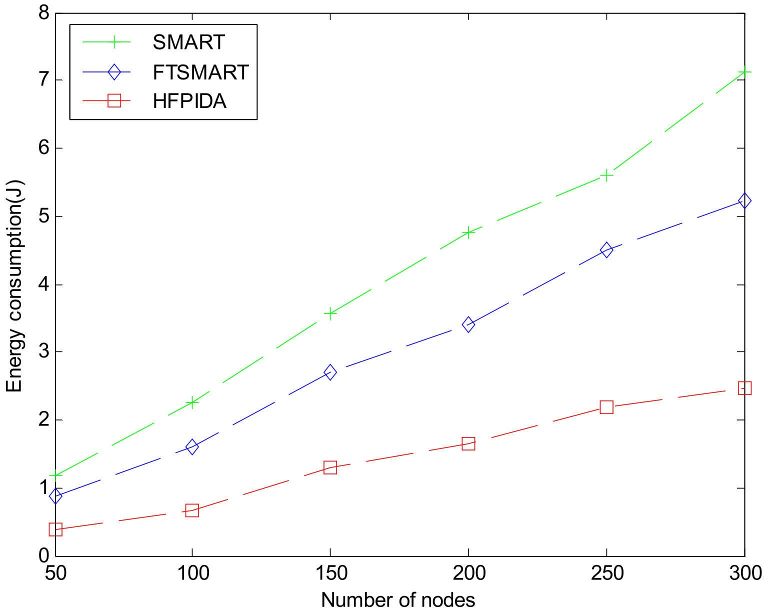

Fig. 3 Energy consumption

the energy consumption of the computation of HFPIDA is lower than that of SMART and FTSMART.

Figure 3 demonstrates the total energy consumption of the HFPIDA, SMART and FTSMART under different number of nodes. It can be observed from the Fig. 3 that, as the number of the nodes increases, the energy consumption of the three schemes increases, too. However, the energy consumption of the SMART and FTSMART is higher than HFPIDA, that is because each node needs to send slices to other neighbor nodes or its parents, there are more messages exchanges for each node in the SMART and FTSMART, and the energy consumption of the computation in the SMART and FTSMART is higher than HFPIDA, too.

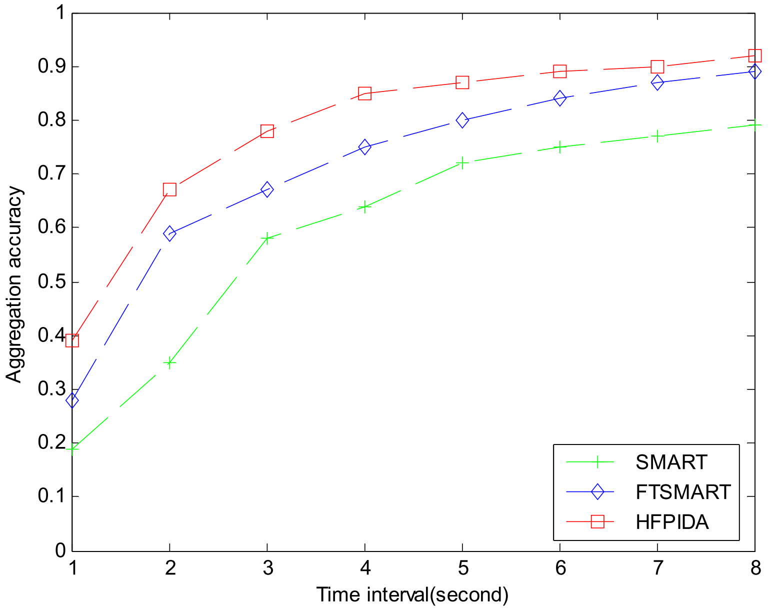

Fig. 4 Aggregation accuracy

Aggregation accuracy

The aggregation accuracy is another important metric to demonstrate the performance of the data aggregation scheme, due to packet losses, delays, collisions and noisy communication channels frequently occur in wireless sensor networks, the accuracy of the aggregation result does not achieve 100%. The aggregation accuracy is given by

$$P_{AC} = \frac{D}{D_t} \tag{29}$$

Where P_{AC} denotes the aggregation accuracy, D denotes the final aggregation result obtained by the base station, D_t denotes the sum data of all nodes in whole network.

Figure 4 shows the aggregation accuracy of the HFPIDA, SMART and FTSMART under different time interval. It can be observed from the Fig. 4 that the aggregation accuracy increases as the time interval increases. That is because the packets have less chance to collide with the longer time interval. It can be observed from the Fig. 4 that the aggregation accuracy of the HFPIDA is the highest, and the aggregation accuracy of the SMART is the lowest. That is because the communication overhead of HFPIDA is the lowest, the communication overhead of SMART is the highest, the more packet transmitted, the more the probability of collision during the aggregation, the more packet lost, which greatly affect aggregation accuracy.

Conclusion

In the process of data aggregation in wireless sensor networks, it is a challenging task to meet both data privacy protection and data integrity verification. In order to protect data privacy and verify data integrity, moreover, balance the energy consumption and security during the data aggregation, a privacy and integrity–preserving data aggregation scheme for wireless sensor networks based on homomorphic fingerprinting (HFPIDA) is proposed in this paper. In the HFPIDA, it only uses lightweight homomorphic fingerprint technology and privacy function, and does not produce any redundant data. Security analysis demonstrates that the HFPIDA is efficient to resist various passive and active attacks launched by attackers, and protects the data privacy and data integrity. Simulation results show that The HFPIDA requires less communication and energy overheads, and can improve the data aggregation accuracy. In the future, the researches on supporting multi-parameters data aggregation and the security protection of multi-parameters data aggregation for wireless sensor networks will be huge challenges.

Authors' contributions

Zhiming Zhang proposed the research topic and designed research plan. In addition, he is also responsible for designing the framework of the manuscript, drafting the manuscript, revising the manuscript and completing the manuscript. Wei Yang plays an important role in researching and organizing literatures, writing and revising manuscript. Fuying Wu is mainly responsible for experimental design and data analysis. Ping Li designs computer programs, executes computer programs and collects data. All authors read and approved the final manuscript.

Funding
This work was supported in part by the National Natural Science Foundation of China under Grant 62002143, and in part by the Science and Technology Research Project of Jiangxi Provincial Department of Education under Grant GJJ2200334.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 26 July 2023 Accepted: 27 September 2023
Published online: 07 October 2023

References

1. Muduli L, Mishra DP, Jana PK (2018) Application of wireless sensor network for environmental monitoring in underground coal mines: A systematic review. J Netw Comput Appl 106:48–67
2. Aslan YE, Korpeoglu I, Ulusoy Ö (2012) A framework for use of wireless sensor networks in forest fire detection and monitoring. Comput Environ Urban Syst 36(6):614–625
3. Liyanage M, Braeken A, Kumar P, Ylianttila M (2020) IoT Security: Advances in Authentication. John Wiley & Sons, Hoboken, NJ
4. Dhanvijay MM, Patil SC (2019) Internet of things: A survey of enabling technologies in healthcare and its applications. Comput Netw 153:113–131
5. Rani S, Maheswar R, Kanagachidambaresan G, Jayarajan P (2020) Integration of WSN and IoT for Smart Cities. Springer, Berlin
6. Grover J, Sharma S (2016) Security Issues in Wireless Sensor Network - A Review, 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Amity University, Noida, India, 2016, pp. 397–404
7. Sert SA, Fung C, George R, et al (2017) An efficient fuzzy path selection approach to mitigate selective forwarding attacks in wireless sensor networks. IEEE International Conference on Fuzzy Systems, Naples, Italy, 2017, pp. 1–6
8. Ozdemir S, Çam H (2009) Integration of false data detection with data aggregation and confidential transmission in wireless sensor networks. IEEE/ACM Trans Netw 18(3):736–749
9. Lakshmi V, Deepthi P (2019) A secure channel code-based scheme for privacy preserving data aggregation in wireless sensor networks. Int J Commun Syst 32(1):1–21
10. He W, Liu X, Nguyen H, Nahrstedt K, Abdelzaher T (2007) PDA: privacy-preserving data aggregation in wireless sensor networks. Proc 26th IEEE International Conference on Computer Communications. IEEE Press, Anchorage, AK, USA, 2007, pp. 2045–2053
11. Li C, Zhang G, Mao Y, Zhao X (2021) A data Aggregation privacy protection algorithm based on fat tree in wireless sensor networks. Security and Communication Networks 2021(8):1–9
12. Alghamdi WY, Wu H, Kanhere SS, Reliable and secure end-to-end data aggregation using secret sharing in wsns. (2017) IEEE Wireless Communications and Networking Conference (WCNC). San Francisco, CA, USA 2017:1–6
13. Hua P, Liu X, Yu J, Dang N, Zhang X (2018) Energy-efficient adaptive slice-based secure data aggregation scheme in WSN. Procedia Comput Sci 129:188–193
14. Zhou L, Ge C, Hu S, Su C (2019) Energy-efficient and privacy-preserving data aggregation algorithm for wireless sensor networks. IEEE Internet Things J 7(5):3948–3957
15. Fang W, Wen X, Xu J, Zhu J (2019) CSDA: a novel cluster-based secure data aggregation scheme for WSNs. Cluster Comput 22(3):5233–5244
16. Elhoseny M, Yuan X, El-Minir HK, Riad AM (2016) An energy efficient encryption method for secure dynamic WSN. Security and Communication Networks 9(13):2024–2031
17. Elhoseny M, Elminir H, Riad A, Yuan X (2016) A secure data routing schema for WSN using elliptic curve cryptography and homomorphic encryption. Journal of King Saud University-Computer and Information Sciences 28(3):262–275
18. Dener M (2022) SDA-RDOS: A New Secure Data Aggregation Protocol for Wireless Sensor Networks in IoT Resistant to DOS Attacks. Electronics 11(24):1–30
19. Parmar P, Kadhiwala B (2016) Secure data aggregation protocol using AES in wireless sensor network. Emerging Research in Computing, Information, Communication and Applications. Springer, Singapore, 2016, pp. 421–432
20. Boubiche DE, Boubiche S, Toral-Cruz H, Pathan A-SK, Bilami A, Athmani S (2016) SDAW: secure data aggregation watermarking-based scheme in homogeneous WSNs. Telecommun Syst 62(2):277–288
21. Liu X, Zhang X, Yu J, Fu C (2020) Query privacy preserving for data aggregation in wireless sensor networks. Wirel Commun Mob Comput 2020:1–10
22. Goyal N, Dave M, Verma AK (2020) SAPDA: Secure authentication with protected data aggregation scheme for improving QoS in scalable and survivable UWSNs. Wireless Pers Commun 113(3):1–15
23. Chenthil TR, Jayarin PJ (2022) An Energy Aware Multi Slot Scheduling with Two-Layer Hexagonal Based Integrated Aggregation Approach for Underwater Wireless Sensor Networks (UWSN). J Interconnection Netw 22(4):44–71
24. Ozdemir S, Peng M, Xiao Y (2015) PRDA: polynomial regression-based privacy-preserving data aggregation for wireless sensor networks. Wirel Commun Mob Comput 15(4):615–628
25. Sreenivasulu AL, Chenna RP (2020) NLDA non-linear regression model for preserving data privacy in wireless sensor networks. Digital Communications and Networks 6(1):101–107
26. Hendricks J, Ganger GR, Reiter MK (2007) Verifying Distributed Erasure-Coded Data. Proceedings of 26th ACM Symposium on Principles of Distributed Computing, Portland, Oregon, USA, 2007, pp.1–8
27. Low CP, Fang C, Mee J, Ang YH (2007) Load-Balanced Clustering Algorithms for Wireless Sensor Networks. IEEE International Conference on Communications. IEEE, Glasgow, Scotland, 2007, pp.3485–3490.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.