CSDA: a novel cluster-based secure data aggregation scheme for WSNs

Wei Fang[1,2,3] · XueZhi Wen[1] · Jiang Xu[1] · JieZhong Zhu[1]

Received: 11 July 2017 / Revised: 10 September 2017 / Accepted: 13 September 2017 / Published online: 22 September 2017

Abstract With the development of wireless sensor networks, privacy-preserving has become a very important problem in numerous wireless sensor networks (WSN) applications. This paper presents a novel energy-efficient secure data aggregation scheme cluster-based private data aggregation (CSDA) based on cluster privacy-preserving. It has good flexibility and practical applicability using the slice-assemble technology. And, the number of fragments will dynamically change from the change of the network scale. Then, it can reduce communication overhead and energy consumption. Finally, the simulation results show that the proposed aggregation method demonstrates better performance in data aggregation precision, privacy-preserving and communication efficiency than other methods.

Keywords Wireless sensor networks · Privacy-preserving · Data aggregation · Energy efficiency · CSDA

1 Introduction

In recent years, wireless sensor networks (WSNs) have drawn more and more attention in the field of academic research and practical applications. WSNs is a task-based network, which integrates with microelectronics, perception, embedded computing, wireless communication and distributed information processing techniques [1]. These sensor nodes in WSNs complete the tasks of collecting of information, surveillance, and the perceiving environment. During the process of data collection, the single node transmits data to the sink individually which causes large wasting in communication bandwidth and valuable energy resource. Then the information collection work cannot be completed on time and it reduces the efficiency of information collection. Due to the limited resource, the sensor nodes have strong restrictions on the processes of computation, storage, and communication. Data aggregation [2,3] is also applied to wireless sensor networks to alleviate these problems, which is a way for dealing with multiple copies of data or information and integrate the data which are more efficient and more according to users' need by reducing data packets in the network, and then increase the efficiency of information collection. Data aggregation can be considered as a fundamental process to reduce energy consumption and communication overhead to save the limited resources in WSNs [4,5].

Nowadays, WSNs is also an important part of Internet of Things (IOT). The characteristics of openness and self-organization expose its vulnerability to the attackers, which can lead to the loss of its original construction purpose and cause even worse damages [6]. Efficient and feasible secure data aggregation scheme build a firm foundation for the application of WSNs in some important fields such as military, political, economic and so on. How to deal with the above constraints and provide secure data aggregation in WSNs has become an important requirement for the sake of sensitive nature of the sensor data. A lot of efforts have been made to protect WSNs from hostile attacks. Most of the existing methods often have large computational, communi-

✉ Wei Fang
hsfangwei@sina.com

[1] School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, China

[2] Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[3] Department of Electrical & Computer Engineering, University of Florida, Gainesville, USA

cation costs and lower privacy-preserving ability. However, there are some technical challenges needed to be solved, such as provide energy, efficient and secure data aggregation schemes in WSNs. Privacy-preserving in data aggregating has become an effective way to protect the data security in WSNs. Aiming at the secure data aggregation, we propose a new algorithm based on data slicing for the tree-structure network. The energy and communication overhead can be reduced due to fewer data packets transmission. Then, it takes a reasonable control of communication costs and computation overhead.

The rest of the paper is organized as follows. Section 2 discusses some related work. A CSDA scheme for security data aggregation in WSNs in Sect. 3 is presented. Section 4 conducts performance simulation experiments and analyzes the experimental results. Finally, Sect. 5 concludes the paper with a summary .

2 Related work

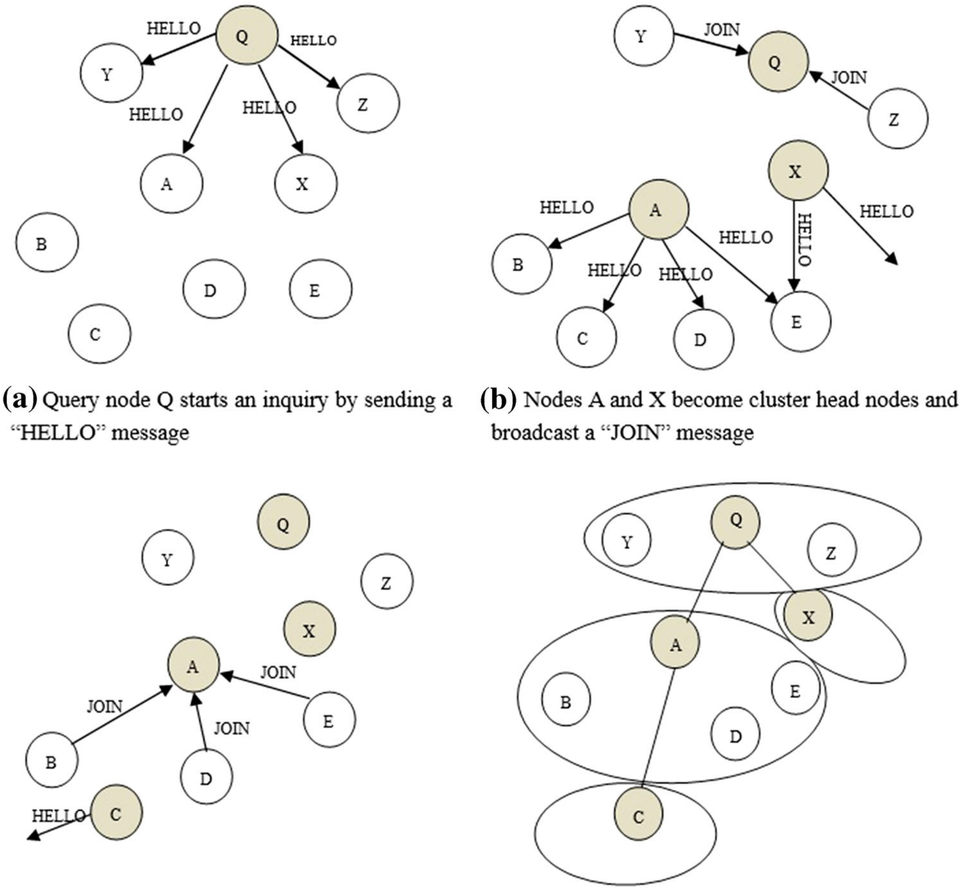
Many effective methods have been addressed by researchers at the data aggregation in WSNs. Rajagopalan and Varshney [2] have given an innovative literature review in the field of data aggregation in WSNs. Jesus et al. [3] review distributed data aggregation algorithms and characterize the different types of aggregation functions. Acharya&all and Armknecht&all, etc. have proposed a CDA(Concealed Data Aggregation) algorithm in which multiple sources nodes sends encrypted data onto a sink along with a convergecast tree [7]. He et al. have put forward a data aggregation algorithm PDA (Privacy-Preserving Data Aggregation), which has two algorithms: CPDA (Cluster-based Private Data Aggregation) and SMART [8]. However, CPDA costs much computational overhead, and SMART also costs too much communication overhead and is sensitive to the loss of data. Over the past few years, several schemes have been proposed in the literature for privacy-preserving data aggregation in WSNs [9–11]. In the field of energy saving, data aggregation reduces the redundant and minimizes the energy consumption for the whole network [12,14].

The security problem of data aggregation in WSNs has become a hot topic. The literature [13–15] gives a summary of security problems on data aggregation in wireless sensor networks. Generally speaking, some basic data aggregation technology cannot provide good data privacy preserving mechanisms. However, in the real world, privacy preserving mechanism is essential and important. For example, in a health monitoring system, sensor nodes can get access to data of patients' vital signs, such as temperature, blood pressure, pulse and so on [16]. These data belong to personal privacy, and patients do not want them to be let out. The secure data aggregation technology in WSNs is that under the circumstances of ensuring the accuracy of data aggregation results, regardless of transferred data being captured and decrypted by external or other internal credible nodes, and can prevent capturing private data. Traditional private data aggregation technology includes CPDA, cluster-based private data aggregation technology, and SMART, slice-assemble data aggregation technology, and some better algorithms, such as k indistinguishable approach of private data aggregation and so on. Several previous methods all get aggregation results in the case of not disclosing private values of any other nodes. The k indistinguishable algorithm of private data aggregation uses a falsity set to make data fuzzy, in which way the real value cannot be distinguished from other k-1 data, instead of encryption technology. This method costs less energy resource than end-to-end encryption and is more efficient than hop-by-hop encryption [17–21].

The rapid development of network technology and its evolution of heterogeneous networks have increased the demand to support automatic monitoring and the management of heterogeneous WSNs [22]. Zhang et al. [23] research how mobile sensors can be efficiently relocated to achieve k-barrier coverage. And, two important problems are studied: relocation of sensors with a minimum number of mobile sensors and formation of k-barrier coverage with minimum energy cost. They are formulated as 0–1 integer linear programming (ILP). Then, they try to solve the relaxed 0–1 ILP rapidly through linear programming. During the process of data aggregation, many aggregation methods are part of the loss data aggregation. They reduce the amount of data transmission by leaving out some details or lowering data quality to save network energy. Therefore, data aggregation operation is also faced a lot of security threats, such as data tampering, data falsification, data discarding and so on, which make users not access to the accurate and complete information.

In the case, the CPDA algorithm can ensure the accurate degree of data, and prevent external nodes to obtain data privacy, but the interaction of the cluster nodes transfers increases the communication overhead of network, at the same time polynomial arithmetic and the inverse matrix calculation also brings a lot of overhead, so that this algorithm application has great limitation. To sum up, above existing schemes fail to make full use of the natural advantage of data link and have higher communication and computational costs. Therefore, this paper has improved the CPDA and proposed a lightweight security data aggregation scheme: CSDA, which uses the approach fragment reassembly and can dynamically adjust the node numbers of the cluster partitions according to the size scale and numbers of WSNs. It meets the general privacy protection requirement, and not only has better flexibility and performance of data aggrega-

Fig. 1 The formation of cluster



(a) Query node Q starts an inquiry by sending a "HELLO" message

(b) Nodes A and X become cluster head nodes and broadcast a "JOIN" message

(c) Node E chooses one cluster to join in randomly

(d) Finally, four clusters formation

tion precision but also costs less communication overhead and lower energy consumption.

3 The CSDA scheme for data aggregation in WSNs

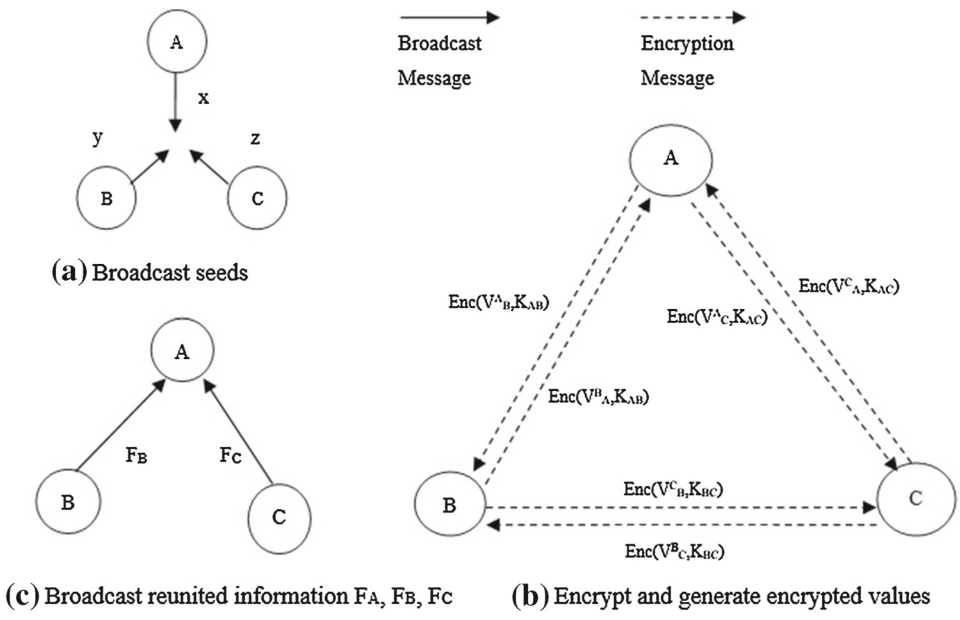
3.1 Introduction of CPDA

The cluster-based private data aggregation (CPDA) algorithm using the fragmentation and reassembly technology can adjust the cluster nodes slice according to the number of the size of a dynamic network scale and has better flexibility and practicability. It comprises three stages: the first stage is cluster formation, the second stage is intra-cluster data aggregation, and the third stage is data aggregation between clusters.

(1) The cluster formation is as shown in Fig. 1. Firstly, query node Q triggers the initial inquiry of a cluster by sending a message "HELLO". When the node receives the message, it decides itself whether to be a cluster head node according to the probability pc, which is a parameter that all nodes have chosen in advance. If one node decides itself to be a cluster head node, it will transmit the message "HELLO" to its neighbor nodes. Otherwise, it will broadcast a message "JOIN" to suggest that it wants to join in a cluster.

In Fig. 1a, node Q starts a query. In Fig. 1b, once nodes Y and Z decide to join in a cluster whose head node is node Q, they will send a message "JOIN" to node Q. Meanwhile, the nodes A and X both decide to be cluster head nodes, and they transmit "HELLO" to their neighbor nodes, hoping that their own clusters come into being. As shown in Fig.1, node E can receive two messages "HELLO". In Fig. 1c, repeat the above-mentioned process and node E chooses one cluster head node between A and X in random. Finally, as shown in Fig. 1d, there will be four clusters with the nodes Q, A, X, C as cluster head nodes in their own clusters.

(2) Data calculation in one cluster. To describe succinctly, we choose one cluster with three nodes, in which A is the head node, nodes B and C are member nodes. And a, b, c represent private data of the three nodes respectively. This privacy protection algorithm is based on the addition property of polynomial. Figure 2 shows the process of calculated data interchange in a cluster during which we can get the sum



(a) Broadcast seeds

(c) Broadcast reunited information F_A, F_B, F_C

(b) Encrypt and generate encrypted values

Fig. 2 Encrypted data exchanging process in one cluster

of private data without letting out private data. At first, as shown in Fig. 2a, every node in the cluster shares a non-zero number with any other node and becomes a seed. We suppose that the seeds of A, B, C are respectively x, y, z and they are different from each other. Then, as shown in Fig. 2b, node A generates two random numbers r_1^A and r_2^A, and the two numbers are only known by the node A. Similarly, node B generates two random numbers r_1^B and r_2^B, and node C generates two random numbers r_1^C and r_2^C.

Therefore, node A can work out three numeric values as shown in (1)

$$\begin{aligned} \mathrm{v}_A^A &= a + r_1^A x + r_2^A x^2, \\ \mathrm{v}_B^A &= a + r_1^A y + r_2^A y,^2 \\ \mathrm{v}_C^A &= a + r_1^A z + r_2^A z^2 \end{aligned} \tag{1}$$

Similarly, node B can also get three numeric values as shown in (2):

$$\begin{aligned} \mathrm{v}_A^B &= b + r_1^B x + r_2^B x^2, \\ \mathrm{v}_B^B &= b + r_1^B y + r_2^B y^2, \\ \mathrm{v}_C^B &= b + r_1^B z + r_2^B z^2 \end{aligned} \tag{2}$$

Likewise, node C may be calculated to get three numeric values in (3):

$$\begin{aligned} \mathrm{v}_A^C &= c + r_1^C x + r_2^C x^2, \\ \mathrm{v}_B^C &= c + r_1^C y + r_2^C y^2, \\ \mathrm{v}_C^C &= c + r_1^C \mathrm{z} + r_2^C z^2 \end{aligned} \tag{3}$$

Then, A uses the encryption key shared with B to encrypt v_B^A and v_C^A, then transmit them to B and C respectively. In the same way, B sends encrypted v_A^B and v_C^B to A and C respectively, and the node C sends encrypted v_A^C and v_B^C to A and B respectively. As a result, we can finally get three values at the node A: v_A^A, v_A^B and v_A^C, then the value of $\mathrm{F_A}$ can be obtained by calculation formula, $\mathrm{F_A} = \mathrm{v}_A^A + v_A^B + v_A^C = (a+b+c) + r_1 x + r_2 x^2$, where $r_1 = r_1^A + r_1^B + r_1^C$ and $r_2 = r_2^A + r_2^B + r_2^C$. Similarly, the value of $\mathrm{F_B}$ can be got from node B by calculation formula, $\mathrm{F_B} = \mathrm{v}_B^A + v_B^B + v_B^C = (a+b+c) + r_1 y + r_2 y^2$, and the value of $\mathrm{F_C}$ can be computed at the node C by calculation formula, $\mathrm{F_C} = \mathrm{v}_C^A + v_C^B + v_C^C = (a+b+c) + r_1 z + r_2 z^2$. And then the nodes B and C transmit the two values $\mathrm{F_B}$ and $\mathrm{F_C}$ to the node A, as we can see from Fig. 2c. And the following values below can be got at the node A as shown in (4):

$$\begin{aligned} \mathrm{F_A} &= \mathrm{v}_A^A + v_A^B + v_A^C = (a+b+c) + r_1 x + r_2 x^2, \\ \mathrm{F_B} &= \mathrm{v}_B^A + v_B^B + v_B^C = (a+b+c) + r_1 y + r_2 y^2 \\ \mathrm{F_C} &= \mathrm{v}_C^A + v_C^B + v_C^C = (a+b+c) + r_1 z + r_2 z^2 \end{aligned} \tag{4}$$

Now, because node A is aware of the sum of x, y, and z, it can find out the result of data aggregation among A, B and the aggregated values $(a+b+c)$ according to the deformation of the calculation formula (5),

$$\begin{aligned} &\mathrm{U} = \mathrm{G}^{-1} F. \\ &G = \begin{bmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{bmatrix}, U = \begin{bmatrix} a+b+c \\ r_1 \\ r_2 \end{bmatrix}, \\ &F = [\,F_A \; F_B \; F_{\mathbf{B}}\,]^T. \end{aligned} \tag{5}$$

(3) The data aggregation between clusters. This step is to deliver clusters' results of data aggregation to the nodes that

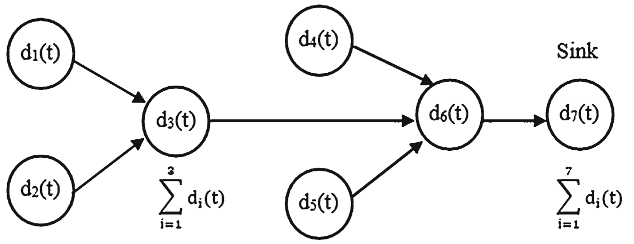


Fig. 3 Schematic diagram of data aggregation

start a query initially by logical topologies. We suppose that logical topologies are not provided with effective in privacy protection, and we mainly discuss the effectiveness in the privacy of the first two steps.

3.2 CSDA: modification of the CPDA scheme

In this section, we provide a new CSDA scheme based on the algorithm CPDA, which has more efficient and secure. Under the general data aggregation of privacy protection, the goal of CSDA scheme is to minimize the data communication and computation among nodes and finally can obtain more precise data aggregation result.

3.2.1 Network model

This sensor network is modeled as a connected graph G (V, E), where V is the set of sensor nodes, $|V| = N$ is the number of sensor nodes, and E represents the set of wireless links connecting the sensor nodes. Each sensor node is equipped with a wireless transceiver which can be used to communicate with the sensor nodes within its transmission range.

Sensor nodes in the process of network data aggregation will be divided into three categories: ordinary node, aggregation node, and sink node (Sink). In fact, ordinary node and aggregation node are essentially similar, but their tasks are not same. The ordinary node as a leaf node is only responsible for data collection, and the aggregation node as a non-leaf node of a network is responsible for the collection of data aggregation processing and application of the appropriate function. Sink node is responsible for receiving the final aggregation results and forwarding the query request, and connecting with the network. The data aggregation process of different nodes is shown in Fig. 3.

A data aggregation function is defined as : $y\ (t) = f\ (d_1\ (t) + \mathrm{d}_2(t) + \cdots + d_{\mathrm{n}}(t))$, where d_{i} (t) $(i = 1, 2, \ldots, \mathrm{N})$ is the individual sensor reading at time instant t for node i. The aggregation function is typically a sum, average, max, min, and count, etc. Recently, some researchers focus on the sum function. Since some statistical functions such as count, average, and standard deviation are all based on sum(). To some extent, for nonlinear functions such as max() and min() can also be used to estimate by the sum function and then forward to the sink, a lot of energy will be saved. So, this paper chooses the sum function as our aggregation function, $y\ (t) = \sum_{i=1}^{N} d_i\ (t)$. All the nodes have the sufficient initial energy.

3.2.2 Threat model

The security problem has become more and more important in the field of the wireless sensor networks. The hacker can launch various attacks to undermine the security of data. In this paper, we mainly focus on preventing eavesdropping attacks to protect data privacy in wireless sensor networks.

Threat model includes these threats from untrusted eavesdroppers intercepting or listening to packets, honest but curious 1 nodes in between data transit, and polynomial time adversaries. In attack processing, the attacker has the ability to learn all the communications and obtain user privacy information by monitoring the wireless link. We assume that the attacker can access the security mechanism adopted in wireless sensor networks by capturing a normal node. When an attacker captures the privacy data of a node to other nodes, data privacy will be threatened. Here, data privacy protection in this paper is that data collection from each sensor node besides itself, and any other unknown nodes.

3.2.3 Encryption key distribution

In order to prevent data transmission of nodes from eavesdropping attacks, wireless link transmission data often needs to be encrypted. CSDA adopts the same random key distribution mechanism like SMART and CPDA [24].

(1) A large key pool of K keys and their identities are first generated randomly.

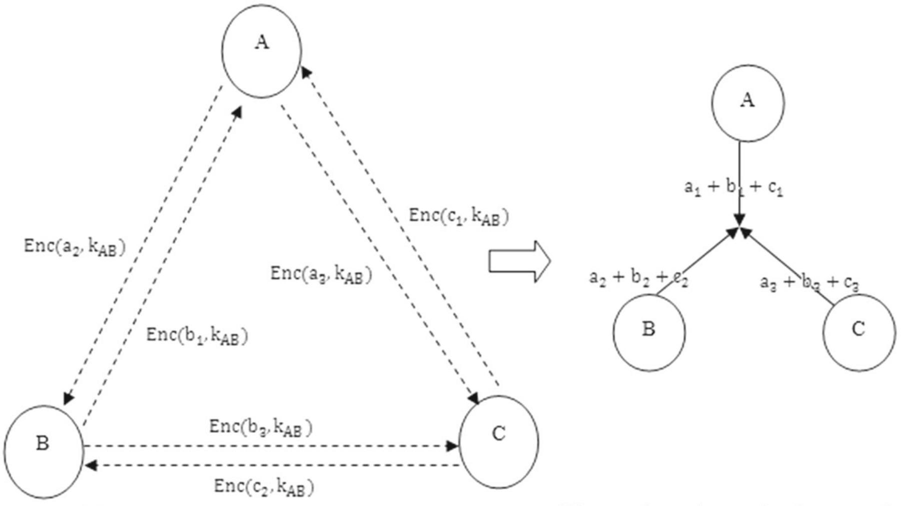
(2) Each sensor node identifies its neighbors which share the same key ($\mathrm{k_i}$) with itself by invoking and exchanging discovery messages. Then, it is possible to establish a secure link between them.

(3) If there are not two neighbor nodes which share the same key, it can be connected by two or more multi-hop secure links.

Therefore, any pair of selected nodes shares the same probability:

$$\mathrm{p}_{\mathrm{connect}} = 1 - \frac{((\mathrm{K}-\mathrm{k})!)^2}{(\mathrm{K}-2\mathrm{k})!\mathrm{K}!} \tag{6}$$

Fig. 4 Data exchanging in the algorithm CSDA



(a) Encrypt and deliver part of values **(b)** Broadcast the result of aggregation

As the third party eavesdropper node which adopts the random key distribution mechanism to select the key-pool of K key randomly, if the K key includes the key_s, then the hacker can use this key_s to overhear the encrypted message, and the probability to obtain key_s is [25]:

$$p_{overhear} = \frac{k}{K} \tag{7}$$

3.2.4 Cluster formation

CSDA is also divided into three steps like CPDA. The first step is the formation of the cluster, which is the same as the first step of the algorithm CPDA. The second step is data aggregation in one cluster. We take advantage of slice-assemble technology applied to it. Assuming that the cluster C_i has m_i members, the node divides its private data into m_i pieces and sends $m_i - 1$ pieces of them to other $m_i - 1$ members. As the Fig. 4 shows, it is the secure multiple data aggregation when there are three nodes.

In Fig. 4a, the three nodes A, B, and C divide their own private data a, b, and c into three slices separately: a_1, a_2, a_3, b_1, b_2, b_3 and c_1, c_2, c_3. They deliver encrypted values as the Fig. 4 shows. And, in Fig. 4b, the node A is to be calculated the value of $a_1 + b_1 + c_1$, and the node B is to be calculated the value of $a_2 + b_2 + c_2$, and the node C is to be calculated the value of $a_3 + b_3 + c_3$. Then they broadcast the three values, and the three nodes add up three values to infer the value of $a + b + c$.

The third step: data aggregation between clusters. In the first two steps of CSDA, the cluster head node is used to data aggregation, while other member nodes are responsible for keeping watch on head node's operations. When the head node of cluster receives intermediate data from head nodes in the downstream direction, it must transmit a message to be known by other nodes in its cluster. For example, as the Fig. 5 shows, the head nodes D and E transfer intermediate results to the node A. Since the nodes B and C are able to pick up the intermediate results from D and E, the two nodes can recognize the correctness of the value which broadcasted by the node A. If the node A cheats on data from D and E, the nodes B and C will check out such a problem and report it to the base station. Then, the aggregation data of cluster Q and A is Q = X + Z, A = B + C + D + E.

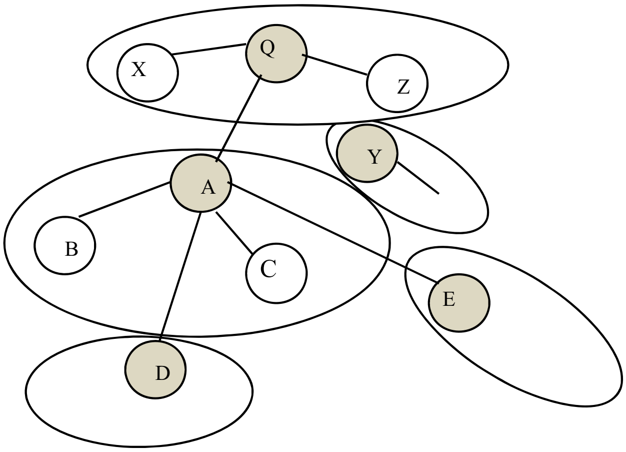


Fig. 5 The cluster and data users

The two head nodes act as data aggregation model may collude with such attacks that can be found to depend on the network topology. Figure 6 shows the circumstance that the nodes X and Y can't detect such a problem, in which X is a head node and Y is a parent node of the head node. In addi-

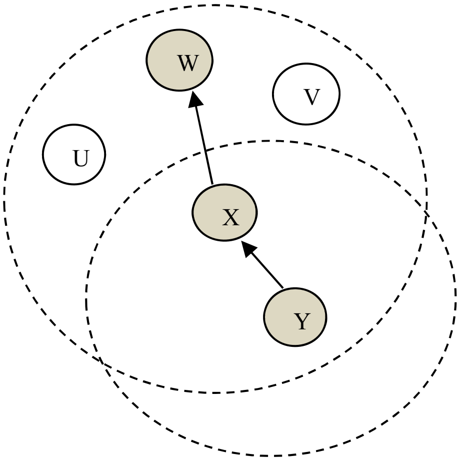


Fig. 6 The monitoring function is out of work

tion, the nodes U, V, and X all belong to the same cluster. If a malicious node exists at the intersection of communication range of X and Y, other nodes will not know the communication content between X and Y, and then the nodes U and V will not judge whether the node X broadcasts the right value or not. However, in a reasonable density network, there is a small probability that such aggressive behavior cannot be discovered. Meanwhile, when the node notices and reports improprieties, it should choose unicast routing and the malicious node which are not in this routing.

3.3 The analysis of effects on privacy protection of CSDA

In the algorithm CSDA, private data can only be revealed when exchanging in one cluster. Supposing that the size of a cluster is m, to any node i in the cluster, only when other $m-1$ nodes have colluded, the private data of the node i will be disclosed.

Aiming to protect data integrity, we analyze the probability of detecting nodes' improprieties successfully, assume that R represents the radius of widespread and the intersection of widespread is no less than $\frac{2}{3}\pi R^2 - \sqrt{3}R^2$ .The average degree of nodes is d. Supposing that friendly nodes are distributed in one area uniformly, we can see from the Fig. 5 that the probability that there are no friendly nodes in the intersection of two areas. It is represented with Eq. (8):

$$\mathrm{P}_{incapable} \le \left(1 - \frac{\frac{2}{3}\pi R^2 - \sqrt{3}R^2}{\pi R^2}\right)^d \approx 0.88^d \qquad (8)$$

In CSDA, d is generally greater than 10. At that time, $P_{incapable}$ is no greater than 0.28; when d is equal to 20, $P_{incapable}$ is no greater than 0.07. That is to say, the higher the density of the network is, the more effective the algorithm is.

3.4 Advantages and disadvantages of CSDA

The CSDA scheme combines the advantages of CPDA and SMART fully. On the one hand, CSDA takes advantage of slice-assemble technology to reduce computational overhead in CPDA algorithm. On the other hand, delivering encrypted data in one cluster reduces data traffic like SMART, and it prevents head nodes from tampering with data. In this paper, an improved algorithm CSDA is proposed to overcome the drawbacks of the SMART algorithm. The SMART algorithm is used to split the tree topology structure of the network, which can cause too much computation and communication overhead. Only leaf nodes of the network will be processed to slices, and send the slice meta information in CSDA. And, non-leaf nodes will only receive the aggregation information. Then, it will greatly reduce the energy consumption of WSNs, and improve the accuracy of aggregation.

Extra loaded in CSDA is the second step and the third step. The head node of a cluster doesn't need broadcast part of aggregation results to members of the cluster. For example, node A does not need to broadcast $a_1 + b_1 + c_1$. But node A broadcasting this value is good for B and C in order to supervise nodes in the third step. Moreover, in the third step, before the cluster head node transfers intermediate results to its parent node, it needs to broadcast the intermediate results it has received. In CSDA, if the density of nodes is appropriate, CSDA will broadcast more $2p_cN$ messages than CPDA, in which N is the number of nodes' data taking part in data aggregation and p_c is the probability that one node chooses itself as a head node.

The CSDA is simulated by MATLAB specific to wireless sensor networks in which the nodes are distributed randomly. The encryption and decryption algorithms of it are custom and simple.

3.5 The algorithm of CSDA

The algorithm description of CSDA is as follows:

```
Input: 1.Collective wireless sensor networks
       2. A SQL type of "summation and integration" query
Output: the results of summation and integration, the number of
messages sent by nodes, calculation times and errors.

The first step: a query of one cluster sponsored by a query node.
    While the number of nodes in an existing cluster increases
        If it receives a query message, it will decide whether it
    becomes the head of the cluster to manage its cluster or the
    normal node in other clusters;
        If it is the head node of the cluster, it will send a query
    message to other nodes that have not been queried yet in its
    communication range to form its own cluster;
    End While

The second step: dealing with every node's data
    For each node in the entire WSN
        Generate its private value ds;
    End For
    For each cluster
        For each node
            Acquire the size of the cluster the node belongs to;
            Generate $size-1$ random values, and divides ds
    into size slices $ds_1$, $ds_1, ds_2 \cdots\cdots ds_{size}$ ;
            In terms of $size-1$ slices of them, it encrypts the
    value of $\text{code} = Encode(key, ds_i)$ and sends it to other
    members in its cluster;
            If the node receives the encrypted value, it will
    decrypt this value, $decode = Decode(key, ds_i)$ and finally,
    sum all values, which is the $sum_i$ of the node.
        End For
    End For
    For each cluster
      For each node in the cluster
        if it is a common node
          It will send $sum_i$ to the head node;
        else
          It will add all the values sent by member nodes to the
           value of itself to get the intermediate result $sum_i$ of the
           cluster;
        end if
      End For
      For each head node of each cluster
          if it is not a query node
         It will send its $sum_i$ to its superior parent nodes ;
        else
         It will sum all the values sent by head nodes and its
          result, compare the final result with the theoretical
          value and calculate errors.
        end if
    End For

The third step: attacking transmitted data to calculate the
performance of this algorithm.
    For each cluster head node
      if the node is on the second layer
        It will expand its sum 100 times secretly.
        if the node has child nodes and there is a friendly node in
the intersection of it and its child nodes
          Child nodes will report errors to the base station;
        else errors will not be discovered;
        end if
      end if
    End For
 The fourth step: the number of loops is enough to get a reasonable
 value of the objective result.
```

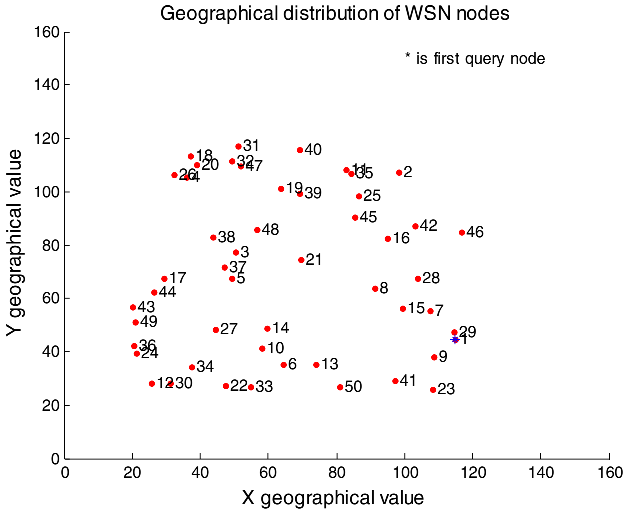


Fig. 7 The WSN topology in CSDA

4 Evaluation and simulations

We mainly compare the effectiveness of CPDA and CSDA in terms of communication overhead and computational overhead through performance simulation experiment.

4.1 Simulation sensor networks

In the simulations, we suppose that fifty sensor nodes come into being randomly in the range of 100 × 100 square meters field. The communication radius is 35 m. The size of data transmitted packet of the network is 40 bytes, and the other network parameters in this paper are not being considered. The cluster sizes range from 3 to 12. And, the maximum routing tree level is 4. In addition, all the algorithms run on the same topology structure of the network in the simulation experiment. Due to the random situation, these computational results are different in every time. In this paper, we analyze and compare the CPDA scheme and CSDA scheme performance mainly from two aspects: privacy preserving capability and communication overhead. The algorithm execution results are shown in the following figures. For example, Fig. 7 shows the topology of a part of clusters.

Figure 7 is the actual geographical distribution of all nodes in wireless sensor networks. Where * represents a query node and its label is 1.

Figure 8 shows the real geographical WSNs child nodes, membership of father nodes. Array A represents that B is father node of A, and is a child node of B.

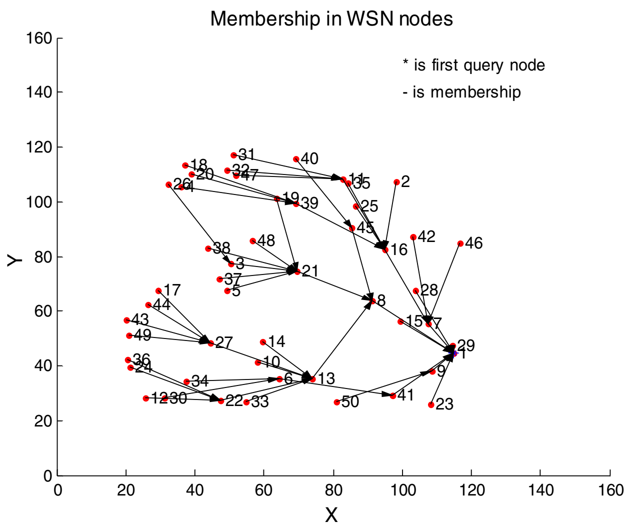


Fig. 8 Membership in WSNs

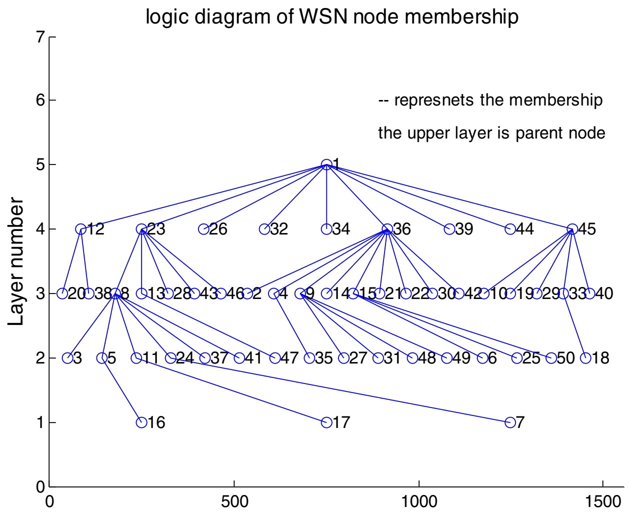


Fig. 9 Clusters topology

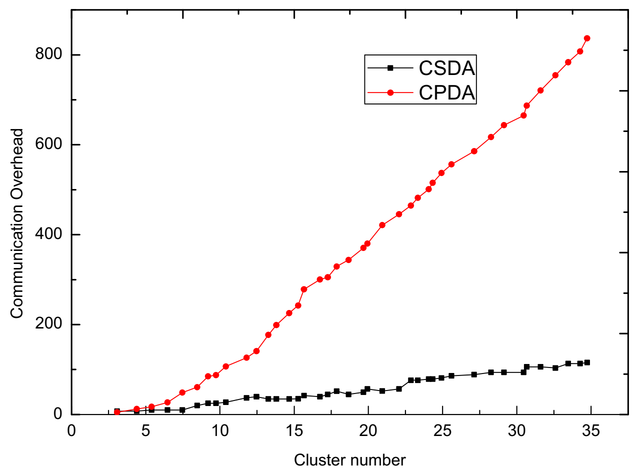


Fig. 10 Communication overhead in the same cluster

Figure 9 shows the logic membership of every node in WSNs, and the upper layer nodes are father nodes, the lower layer nodes are children nodes. For example, the fourth layer node 12 is father node of node 20 and node 38.

4.2 Communication overhead

In order to test the communication overhead of two protocols, the original CPDA protocol, and the proposed modified CSDA protocol), we adopt the total bytes of communication packet during data aggregation as metrics.

While initializing the network, sink node sets each node of the network to a specific time interval (epoch duration), it represents the required time to complete the data aggregation. One "Hello" communication message from each sensor node will generate the aggregation tree, and one message is required for data aggregation processing. When each node of merging tree receives this specific message, it will give its child node an assigned timepiece of Meta message.

In contrast to the communication overhead between CSDA and CPDA from the cluster leader node and 2 message communications from each cluster member node, we utilize the same cluster structure to evaluate them. The experiment result is shown in Fig. 10.

We can see the Fig. 10 that CSDA costs less communication overhead than CPDA in the entire network. The data communication overhead of CPDA generated by nodes in the same cluster with each other is much more than that of CSDA generated by slices divided according to the size of clusters. And due to slice-assemble technology in CSDA, when cluster numbers increase, the increase in data communication overhead is not obvious. But the data communication overhead of CPDA increases largely. If we increase the appropriate number of random distribution nodes which are intended to improve the density of the network, the number of findings will increase. That is to say that the circumstance of supervisory nodes which are not working will reduce the increase of network density. And the higher of network density is, the more efficient the algorithm is. They are in accordance with the theoretical analysis.

The above discussion is mainly about data communication overhead in one cluster. It is very good response quantity of data communication between each cluster, but not consider on the total data communication overhead of the entire network. Next, we compare the data communication overhead of the two schemes in a whole network.

In Fig. 11, we can compare the communication overhead with a different K value between the CSDA algorithm and CPDA algorithm when p is set to the value ($p = 0.3, K = 3$).The X-axis represents epoch duration time of aggregation, the Y-axis represents the amount of data communication

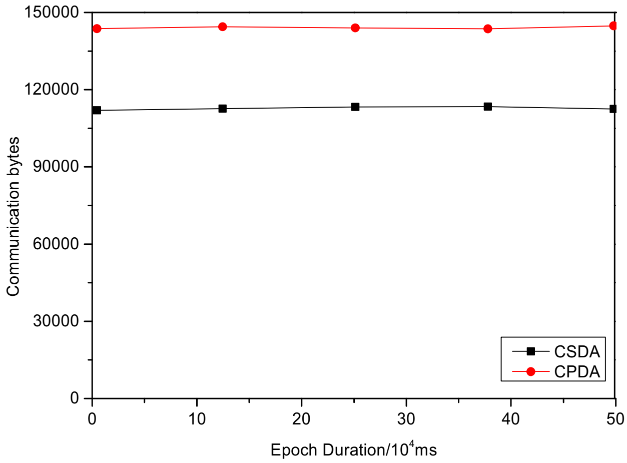


Fig. 11 Communication overhead in a whole network

bytes of aggregation. It can be seen from the Fig. 11 when the K is at the same time. And, the communication overhead of the CSDA algorithm costs lower than that of CPDA algorithm. Since the non-leaf node of CSDA algorithm does not need to send a piece of Meta message to other neighbor nodes, it will also reduce the number of data communication overhead packets in the whole network. In addition, the communication overhead both the CSDA algorithm and CPDA algorithm will increase with the increase of K value. Because when the K value becomes larger, the number of nodes sending out piece will increase linearly, the amount of data communication network will also increase.

4.3 Privacy preserving capability

As we all know, in wireless sensor networks, when the communication channel among nodes eavesdrops or these nodes are mutual cooperation, then the original message of nodes may be leaked. A good data aggregation scheme for privacy-preserving should be able to ensure the privacy message of nodes not to be captured by other points or the hacker.

For each node in two schemes, we represent the cracked probability of measured data as $P_{\text{cdpa}}(q)$ and $P_{\text{csda}}(q)$, where $q$is the cracked probability of links between nodes, and $q \approx P_{\text{overhear}}$.

As CPDA algorithm, each node will be polynomial computed with its own data and non-confidential seed, and then transmit data within the cluster by encryption method. When the cluster size is m, each node needs to transmit $M-1$ encrypted messages to other members of $M-1$ cluster, other nodes only know the $M-1$ encryption key, and the real data of the node can be cracked. Therefore, the cracked probability of an average node using CPDA algorithm is as follows:

$$P_{\text{cpda}}(q) = \sum_{k=m_{\text{c}}}^{d_{ma}\text{x}} P(m=k)\,(1-\left(1-q^{k-1}\right)^{k}) \qquad (9)$$

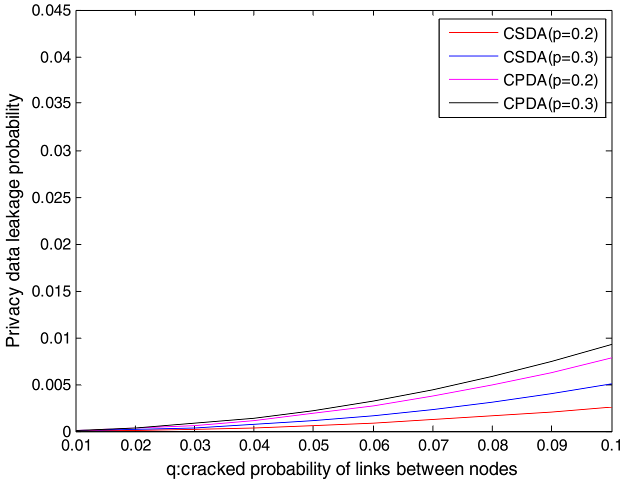


Fig. 12 Privacy preserving performance comparison between CSDA and CPDA

where $d_{\max}$ represents the maximum node number of cluster, m_{c} is the minimum node number of cluster, $P(m=k)$ represents the probability of the k node number of cluster.

In the CSDA scheme, when outward transferring $(K-1)$ slice value, each node will also receive the slice value. Through the analysis of the attack, the model shows that the hacker wants to steal the original perception data of network node, they must also crack $(K-1)$ link of the node and the number of m in degree links.

$$P_{CSDA}(\text{p}) = q^{k-1}\sum_{\text{m}=0}^{d_{maz}} p\,(in\_degree=m)\,q^{m} \qquad (10)$$

where, $d_{\max}$ represents the maximum in the degree of network nodes, and *p(indegree = m)* represents the probability of node which indegree number is m.

The privacy preserving performance comparison between CSDA and CPDA is shown in Fig. 12. The X axis represents the cracked probability between nodes; the Y axis addresses the privacy data leakage probability. We can see that the smaller p value, the better effect of the privacy protection, and when p is a fixed value, then q will get the worse effect of privacy preserving by greater q value. Since the CSDA utilizes the dynamic slicing method, the slice numbers of cluster nodes can be adjusted with the scale of the network. The $P_{CSDA}(p)$ decreases with the increase of fragmentation, which indicates that privacy protection is increasing gradually. Then, the simulation result shows that the CSDA scheme has better privacy preserving ability than the CPDA scheme.

5 Conclusions

In this paper, we modified the CPDA protocol for WSNs and added the functionality of intrusion detection to secure

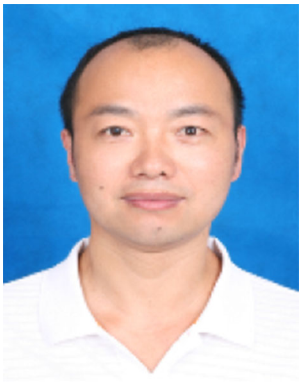
WSNs from the sinkhole, and selective forwarding attacks. The cluster head provides further security protection on the WSNs data. The proposed algorithm based on data slicing using the tree-structure network can reduce the energy and communication overhead. Simulations show that our scheme is feasible, secure and accurate. Next, we plan to evaluate the proposed CSDA schema with other existing privacy-preserving approaches in real wireless sensor networks for data aggregation and provide data integrity protection.

Acknowledgements This work was supported in part by Soft Science Program of China Meteorological Administration (2017[29]), the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey [J]. Comput. Netw. **38**(4), 393–422 (2002)
2. Rajagopalan, R., Varshney, P.: Data-aggregation techniques in sensor networks: a survey [J]. IEEE Commun. Surv. Tutor. **8**(4), 48–63 (2006)
3. Jesus, P., Baquero, C., Almeida, P.: A survey of distributed data aggregation algorithms [J]. IEEE Commun. Surv. Tutor. **17**(1), 381–404 (2015)
4. Qayyum, B., Saeed, M., Roberts, J.: Data aggregation in wireless sensor networks with minimum delay and minimum use of energy: a comparative study [C]. In: Mobility, Intelligent Networks and Smart Societies, pp. 573–582 (2015)
5. Randhawa, S., Jain, S.: Data aggregation in wireless sensor networks: previous research, current status and future directions [J]. Wirel. Pers. Commun. **4**, 1–71 (2017)
6. Acharya, M., Girao, J., Westhoff, D.: Secure comparison of encrypted data in wireless sensor networks [C]. In: Proceedings of the 3rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WIOPT), Washington, DC, USA, pp. 47–53 (2005)
7. Armknecht, F., Westhoff, D., Girao, J., Hessler, A.: A lifetime-optimized end to-end encryption scheme for sensor networks allowing in-network processing [J]. Comput. Commun. **31**(4), 734–749 (2008)
8. He, W.B, Liu, X., Hoang, N., et al.: PDA: Privacy-Preserving Data Aggregation in Wireless Sensor Networks [C]. In: IEEE INFOCOM'07, vol. 28, pp. 2045–2053. IEEE Press (2007)
9. Fontaine, C., Galand, F.: A survey of homomorphic encryption for nonspecialists [J]. In: EURASIP J. Inf. Sec. 2007(1), 1–10 (2007)
10. Sen J., Maitra S.: An Attack on Privacy-Preserving Data Aggregation Protocol for Wireless Sensor Networks [C]. In: NordSec 2011 LNCS 7161, pp. 205–222 (2012)
11. Perrig, A., Stankovic, J., Wagner, D.: Security in wireless sensor networks [J]. Commun. ACM **47**(6), 53–57 (2004)
12. Rajalakshmi, M.C., Gnana, P.A.P.: REEDA: routing with energy efficiency data aggregation in wireless sensor network [C]. In: International Conference on Emerging Research in Electronics, Computer Science and Technology. pp. 174–179. IEEE (2016)
13. Alzaid, H., Foo, E., Nieto, J.M., Park, D.G.: A taxonomy of secure data aggregation in wireless sensor networks [J]. Int. J. Commun. Netw. Distrib. Syst. **8**(1), 101–148 (2012)
14. Man, D., Wang, C., Yang, W., et al.: Energy-efficient cluster-based privacy data aggregation for wireless sensor networks [J]. J. Tsinghua Univ. **57**(2), 213–219 (2017)
15. Yang, G., Wang, A.Q, Cheng, Z.Y, et al.: An energy-saving-preserving data aggregation algorithm [J]. Chin. J. Comput. **34**(5), 792–799 (2011)
16. Pateriya, R.K., Sharma, S.: The evolution of RFID security and privacy: a research survey [C]. In: 2011 International Conference on Communication Systems and Network Technologies IEEE, pp. 116–117 (2011)
17. Carbunar, B., Yu, Y., Shi, L., et al.: Query privacy in wireless sensor networks [C]. In: 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON '07), pp. 203-212. IEEE Computer Society (2007)
18. Kur, J., Stetsko, A.: Location Privacy Issues in Wireless Sensor Networks [J]. The Future of Identity. In: IFIP AICT vol. 298, pp. 160–169 (2009)
19. Chen, J., Zhang, H.L.: Survey on wireless sensor network security [J]. J. Harbin Inst. Technol. **43**(7), 90–95 (2011)
20. He, W., Nguyen, H., Liu, X., Nahrstedt, K., AbdeIzaher, T.: iPDA: an integrity-protecting private data aggregation scheme for wireless sensor networks[C]. In: Proceedings of the Military Communications Conference, San Diego, CA, pp. 1–7 (2008)
21. He, W., Liu, X., Nguyen, H., Nahrstedt, K.: A cluster-based protoc01 to enforce integrity and preserve privacy in data aggregation [C]. In: Proceedings of the 29th IEEE International Conference on Distributed Computing Systems Workshops, pp. 14–19. Montreal, QC (2009)
22. Qu, Z.G., Keeney, J., Robitzsch, S., Zaman, F., Wang, X.J.: Multi-level pattern mining architecture for automatic network monitoring in heterogeneous wireless communication networks [J]. China Commun. **13**(7), 108–116 (2016)
23. Zhang, Y., Sun, X., Baowei, W.: Efficient algorithm for k-barrier coverage based on integer linear programming [J]. China Commun. **13**(7), 16–23 (2016)
24. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks [C]. In: Proceedings of the 9th ACM Conference on Computer and Communications Security. Washington, USA, pp. 41–47 (2002)
25. Jaydip, S., Subhamoy, M.: An attack on privacy-preserving data aggregation protocol [C]. Lect. Notes Comput. Sci. **7161**, 205–222 (2012)

Wei Fang is an associate professor in the Jiangsu Engineering Center of Network Monitoring at the Nanjing University of Information Science & Technology in China. His research interests are in the areas of Big Data Mining and Cloud Computing. He is a member of ACM.



XueZhi Wen is an associate professor with the School of Computer and Software, Nanjing University of Information Science and Technology, China. He is a member of ACM. His research interests include Pattern Recognition, Image Processing, and Cloud computing.

JieZhong Zhu is an associate professor with the School of Computer and Software, Nanjing University of Information Science and Technology, China. His research interests mainly include cloud computing, Digital Image Processing.

Jiang Xu has received his M.Sc. at 2003 and is studying for a doctor's degree. Since 2003 has been working as an instructor in the school of Computer and software at the Nanjing University of Information Science and Technology, Jiangsu, China .His research interests include WSNs, Cloud Computing.