Journal of Computers Vol. 27, No. 4, 2016, pp. 171-185
doi:10.3966/199115592016122704014

A New Private-preserving Algorithm Based on SMART

Bowen Han[1], Zhenjiang Zhang[2*], and Lorna Uden[3]

[1] School of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Jiaotong University, Beijing 100044, China
15120074@bjtu.edu.cn

[2] School of Software Engineering, Key Laboratory of Communication and Information Systems, Beijing Jiaotong University, Beijing 100044, China
zhjzhang1@bjtu.edu.cn

[3]School of Computing, FACS, Staffordshire University, The Octagon, Beaconside, Stafford, St180AD. UK.

Received 6 November 2015; Revised 20 January 2016; Accepted 15 February 2016

Abstract. Privacy is a critical issue in many WSN applications because wireless sensor networks (WSNs) are vulnerable to malicious attacks due to their characteristics. Existing hop by hop and shuffling based privacy preserving protocols does not provide an energy efficient, accurate and secure data aggregation because of the energy consuming decryption at the aggregator node. How to provide effective privacy-preserving algorithm during data aggregation is important. Our main aim is to provide an energy efficient and secure data aggregation scheme, which guarantees the privacy, authenticity and freshness of individual sensed data as well as the confidentiality, accuracy and integrity of aggregated data. This paper improved on the limitations of the SMART (Slice-Mix-AggRegaTe) algorithm, and proposes a new privacy protection algorithm for wireless sensor network (WSN). We have different operations between different nodes, bring in probabilistic segmentation, and also add the data coefficient as well as the positive and negative factors. In addition we also consider making exchange time random in the algorithm. Simulation results demonstrate that the proposed algorithm reduces the communication consumption and the data traffic, improves the precision of data aggregation and the degree of privacy-preserving to a certain extent.

Keywords: data aggregation, privacy-preserving algorithm, SMART(Slice-Mix-AggRegaTe)

1 Introduction

Since wireless sensor network (WSN) is widely used in health care, military defense and sensitive sectors such as business, information security and privacy should not be ignored. Due to the unique characteristic of WSN, the privacy protection of WSN is different from the traditional methods applied in the areas of cryptography and information security. For WSN, it is important to understand the characteristics of WSN such as large-scale, distributed, limited resources and data-center, uncertainty and risk of the external environment of nodes. We must also understand that different network environments lead to different security requirements as well as forms of attack. Therefore the traditional privacy-preserving methods are difficult to apply in the WSN, because existing methods mostly borrowed methods used in other fields that have their own shortcomings and different scopes of application.

At the same time, data aggregation must satisfy the security requirements of the applied network, we cannot use safety performance to exchange network efficiency. Although the continuous operation of aggregation has improved the utilization of communication bandwidth and energy efficiency, it has some

* Corresponding Author

unique negative influences on the security of the network. Data aggregation operation must satisfy the security requirements of the whole sensor network. Therefore privacy protection in data aggregation is a problem that must be dealt with urgently [1] during the large-scale use of WSN and Internet of things.

The security threats during aggregation can be divided into internal and external threats. For external threats, the eavesdropper can monitor wireless communication link to steal sensitive information in transmission, and damage nodes during deployment. Internal threats include when attackers trapping or controlling the nodes can get the key to obtain or tamper with the information, or other trusted users using its related key information to crack, revivification the primary data [2]. Researches of privacy protection technology in data fusion have attracted a lot of attention [3] in recent years.

The paper is organized as follows: In the next section we sum up existing concepts and classification of privacy protection technology, analysis and comparisons of different data fusion strategies and summarizes of their advantages and disadvantages as well as the applicable scope. The third section gives a brief introduction on the principle, advantages and disadvantages of SMART (Slice-Mix-AggRegaTe) algorithm. The improved PSPDA algorithm is described in section 4. Simulation, verification and the corresponding comparison with the original algorithm are discussed in section 5. In section 6, we conclude the paper with suggestions for further work.

2 Related Work

2.1 Classification of privacy-preserving algorithm

The general Classification of privacy-preserving algorithm for WSN in general is shown in Fig. 1.

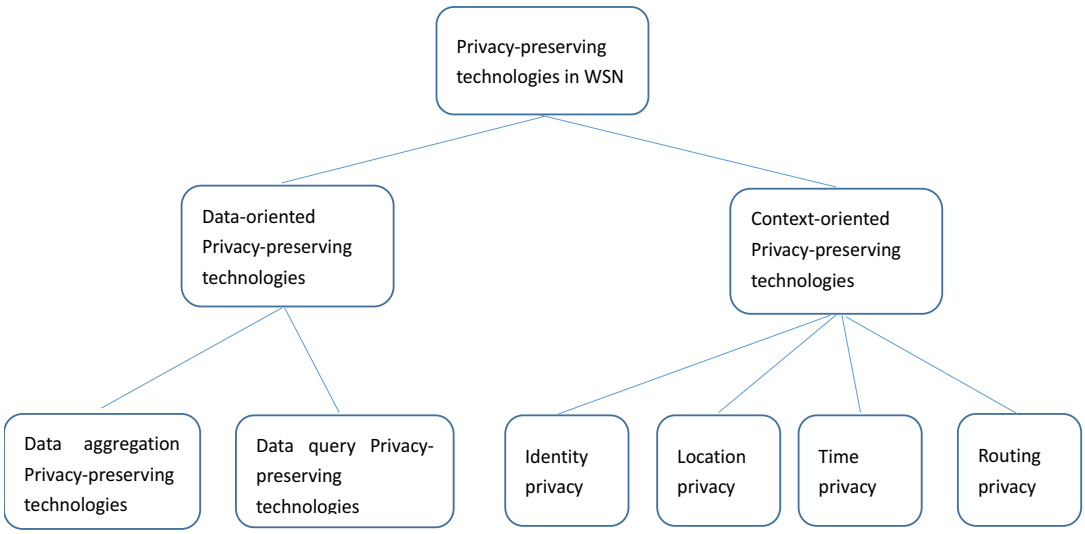

Fig. 1. Classification of privacy-preserving algorithm

As is shown above, the privacy protection algorithms in WSN are roughly classified into Data oriented privacy protection strategy and Context oriented privacy protection. Data oriented privacy protection strategy focuses on the data, including data from the node itself and the data sent from other nodes. It is divided into privacy of data aggregation and data query. Context oriented privacy protection can be divided into identity privacy, location privacy, time privacy and routing privacy [4]. Through these privacy policies, we can encrypt the sensitive information to prevent leakage. Data oriented privacy protection include technologies such as data distortion, anonymous technology and encryption technology and so on. This paper is concerned with the privacy protection of data aggregation to ensure the security of user's data, even if the transfer are intercepted and decoded, the attackers will not be able to restore the user's sensitive data [5].

The existing privacy protection methods are currently divided into: encryption mechanism (including the end-to-end encryption mechanism and hop by hop encryption mechanism), non-encrypted mode, the secure multi-party computation (SMPC) [6] and so on.

The SMPC is one of the privacy protection methods in data aggregation, A. C. Yao first put forward

secure multi-party computation (SMPC) [7] to solve the calculation problems when there is more than one participant and they do not trust each other. It's an algorithm to sum each user's data without letting out the confidentiality and privacy of data. The advantage is that it can guarantee the independence and privacy of information. But there are many disadvantages including computational complexity and communication overhead is too big, and integration approach is limited, which cannot be effectively applied to solve practical problems.

2.2 Hop by hop encryption algorithm

In the hop by hop encryption algorithm [8], two nodes share the same key to effectively cope with external attacks, preventing external enemies to eavesdrop communication links to get plaintext information. In order to prevent internal attack, it uses a data disturbance, segmentation and restructuring or other technologies to prevent controlled nodes from information leakage. Transmissions between nodes also need encryption and decryption operations, to ensure privacy by increasing the computational consumption and delay. He and Liu [9] put forward two kinds of data aggregation privacy-preserving scheme: the CPDA (cluster-based private data aggregation) [10] algorithm—leverages clustering protocol and algebraic properties of polynomials. It has the advantage of incurring less communication overhead**.** The other is SMART algorithm—builds on slicing techniques and the associative property of addition. It has the advantage of incurring less computation overhead.

CPDA is based on clustering protocol and algebraic properties of polynomials. It is divided into three steps: the formation of clusters, data exchange within the cluster nodes, data aggregation of clusters. The more nodes a cluster contains, the more network traffic is needed. The exchange process is accompanied by a large number of calculations, which increases the computational overhead. There is a lack of integrity verification and it also support less aggregation type compared to the SMART algorithm. This must be improved to increase its practicability. The DADPP (data aggregation company privacy-levels protection) [11] algorithm is similar to the CPDA algorithm; it can provide different levels of privacy protection in the data aggregation processing. Different sensor nodes, with different levels of privacy protection, have different groups and numbers. The process is similar to CPDA algorithm. Unlike CPDA algorithm, it preprocess data in the cluster before uploading and the cluster nodes will continue to implement the aggregation operation, and then upload the results to the base station. IPDA [12] and ICPDA algorithm [13] is an improvement on the CPDA. Compared with the CPDA, its advantage is that it is joined to the data integrity authentication mechanism. With the integrity protection strategy: the algorithm is more complex; communication overhead is directly proportional to the increase numbers of nodes; and energy consumption also has a sharp increase, having a serious impact on the performance of WSN and the existence time.

The SMART Algorithm is a kind of hop by hop encryption algorithm. It uses the method of data slicing, mixing and aggregation that provide good privacy protection. However, the traditional SMART algorithm requires all the data sliced to be sent to the adjacent nodes, leading to large data communication in the network, which is prone to crash and resulted in missing data. Hence the disadvantages are: high communication overhead, high energy consumption, low aggregation accuracy and so on.

2.3 End to end encryption algorithm

End to end encryption finishes the aggregation without decrypting. Using this kind of encryption method can complete the aggregation operation on the encrypted data, and the final fusion results in base station is the same with the data fused by hop by hop mechanism, which can effectively respond to external and internal attacks. Non-encrypted mode refers to using other ways such as data distortion, data disturbance or data anonymization to complete the privacy protection in data aggregation. The Advantage is that it can save middle calculation cost for encryption and decryption, reduce time delay, and better able to cope with the internal and external attack at the same time. In the end to end encryption algorithm, because base station and each sensor node share the same key, its use can malicious tampering with privacy information, if nodes are captured by the attacker, these are the disadvantages. AHE (additive homomorphic encryption) [14] algorithm is a kind of privacy homomorphism representative algorithms. It is based on the current key mechanism; node and base station no longer share the same key. Because keys are relevant to ID on those nodes, therefore AHE algorithm has good performance of privacy protection. It

requires uploading each ID's information thereby resulting in increasing communication overhead and energy consumption. AHE algorithm also does not support integrity verification. There is also CDA (concealed data aggregation) [15] algorithm based on privacy homomorphism encryption of data aggregation strategy. It applies the homomorphic encryption mode proposed by Domingo Ferrer (DF scheme) to complete privacy protection in the process of data aggregation in WSN. The algorithm supports multiple fusion operation; at the same time it also applied the segmentation techniques to prevent internal attacks and homomorphic encryption to prevent external attack. IPHCDA (integrity protecting hierarchial concealed data aggregation) [16] make an improvement on CDA algorithm, adding the integrity verification and applying the homomorphic encryption scheme based on elliptic curve [17] and message authentication codes. At the same time, because CDA algorithm cannot provide integrity verification, the base station cannot get the original information. RCDA algorithm is put forward to address this problem [18]. It not only perform data recovery, when compared with the CDA algorithm, it reduces computational complexity, network traffic and greatly improve the network performance.

Although the above methods can be used in privacy protection during data aggregation, each has limitation. This paper proposes a new algorithm to overcome the limitations of SMART. The new algorithm is useful for WSNs with high safety, lower energy consumption and higher precision. It can also ensure reduction in communication overhead and the probability of conflicts by making the node types, number and size of slices, the exchanging time of slices different. At the same time, it will improve the fusion accuracy by introducing data collusion coefficient. This new algorithm retains the good features of the original SMART algorithm and overcomes many of its limitations.

3 An introduction to SMART algorithm

The improved algorithm proposed by this paper is based on SMART algorithm, here is a brief review of SMART.

3.1 The principles of the SMART algorithm

Subdivision is the basic rule of the SMART (Slice-Mix-AggRegaTe) [19] algorithm. It applies the segmentation and recombinant technology. Slicing, mixing and aggregation are the three most important steps. In SMART we should first set the maximum slices number of each node and the collected data will be sliced and randomly assigned to different nodes within h hops, with one slice keeping in local. We usually take h for 1, and each node then mix with all the slices including the received slices, upload aggregated data step by step according to the tree structure previously built. It cannot be restored unless the attacker can obtain all the slices, ensuring the confidentiality of data.

SMART algorithm is divided into three stages:

Step1: slicing: first we set a maximum slice number J for nodes; each node randomly selects a set of nodes within h hops as the destination node set. Usually, we take h =1. Then node slices its private data into J pieces, one of which keep at itself while others encrypted and sent to the set (See Fig 2).

Step 2: mixing: When a node receives the encrypted slice, it decrypts the data using the shared key, and then mixes and sums up all the pieces of slices to get a new data and hide the original information (See Fig 3).

Step 3: Aggregation: data in nodes are aggregated and the results are sent to the base station, finally the base station get the result of aggregation (See Fig 4).

Assuming that $J = 3$, SMART algorithm can be described as follows: d_{ij} represents the slice from $node_i$ will be sent to $node_j$, while the dii means the local slice.

3.2 Advantages and disadvantages of the SMART algorithm

The four characteristics of SMART are summarized below:

High communication overhead and energy consumption. In SMART, each node slices the collected data into J slices and assign them to the adjacent nodes within h jump, thus increasing the number of data to be exchanged in a network. It also leads to a huge increase in communication load as well as energy consumption. Because data traffic is proportional to the largest subdivision number, the statuses of all nodes are the same in the SMART, so the number of packets each node should send is $J - 1$. Energy con

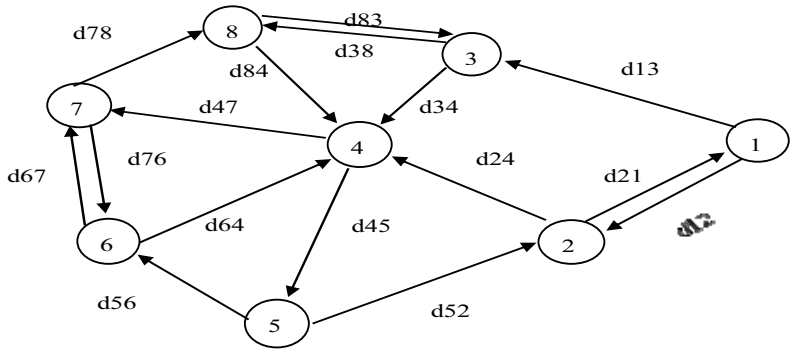

Fig. 2. Slicing

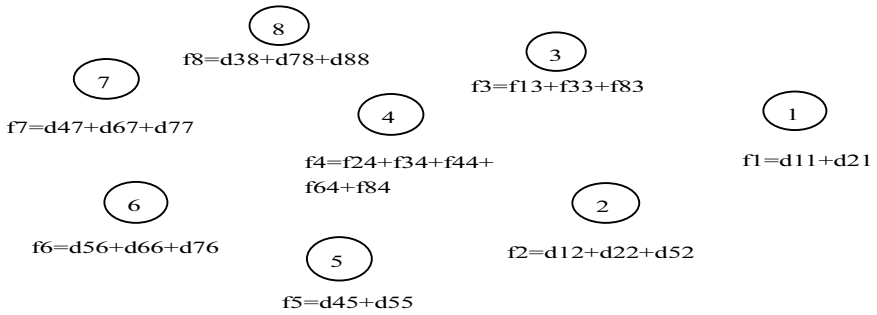

Fig. 3. Mixing

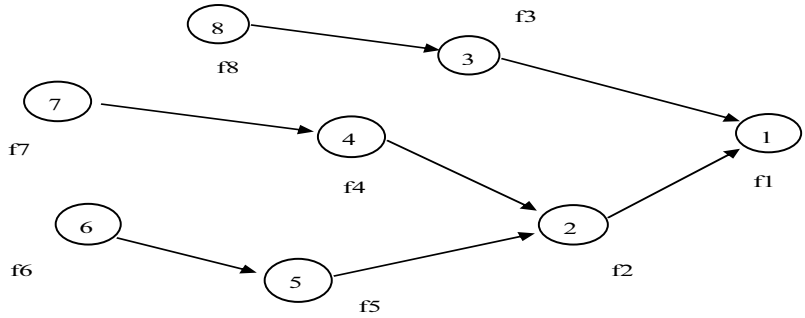

Fig. 4. Aggregation

sumption generally contains the communication overhead and the computational overhead, because the communication and transport costs make up most of the energy consumption. This will affect the performance of WSN network due to greatly increased communication cost in SMART.

High performance of privacy protection. SMART algorithm can prevent from internal threats by slicing thus making the data not exposed to the parent node. This is still not adequate even if the data are exposed to the parent node; while the hop by hop encryption mode and random key distribution scheme cope effectively with the external attack. Through the analysis we know that the bigger the J, the more the number of slices. It is unlikely for an attacker to obtain all slices, which is to say, the better the privacy protection performance of SMART. However, if J is too big, it may cause communication overhead, so we should take a proper subdivision number J to balance traffic and privacy protection performance. Generally the leakage rate of sensitive information is less than 0.5% when J is bigger than 3, which is enough for common security requirements. So we take $J = 3$.

Low accuracy of aggregation. Ideally, the data fusion precision should be 1, but the data collected by node will be sliced, and then randomly assigned to the adjacent nodes. When each node starts to exchange the slice, there are a lot of data packets to be sent thus increases the probability loss data. . When combined with the reality of the network environment, this leading to the amount of packets received in

base station less than the original total amount. So the fusion precision is less than 1. The aggregation accuracy of SMART is not high and it needs to be further improved.

Lack of integrity verification. There are no integrity protection mechanisms in SMART. Base station can't find the data that have already changed if attackers maliciously tamper with the data or add the wrong data.

As can be seen from the above summary, SMART algorithm improves the performance of privacy protection, but in terms of energy consumption, network traffic, and integrity verification, it needs to be further strengthened. Otherwise, its uses in real application are very limited. .

This paper proposes a new algorithm that overcomes the limitations of SMART for WSN.

4 The Improved PSPDA Algorithm

4.1 The proposed improved algorithm

This section introduces the improved algorithm PSPDA (Particular-Sliced Private-preserving Data Aggregation) adopted from the SMART algorithm. The proposed algorithm suggests a series of improvement to SMART by making it a more suitable algorithm with low energy consumption, high privacy protection performance and high precision of data aggregation for WSN.

In this paper, WSN network is a connected graph composed of the node S_i and the link L_i, the vertex S_i ($S_i \in S$) is on behalf of the WSN nodes within the network and S is the set of sensor nodes, while the L_i, represents the link between adjacent nodes and L is the communication link set. *NodeNums* stands for the total number of nodes in WSN.

There are three types of nodes in our proposed WSN models: the BS, the aggregation nodes and ordinary leaf nodes. In order to facilitate later discussion, we assume that there only exists one base station in the network center. During aggregation, common leaf nodes only responsible for collecting data and then aggregate the mixed slices and send it to the aggregation nodes; Base station node is equal to the task management terminal part of the sensor network and data fusion results are sent to the base station.

This paper only discusses the typical sum aggregation function; we define the fusion function as

$$f(t) = \sum_{i=1}^{n} f(d_i(t)) \quad \textbf{(1)}$$

where $d_i(t)$ (i =1,2….n) represents the data collected in t

The schematic diagram of the sum aggregation function is as Fig. 5:

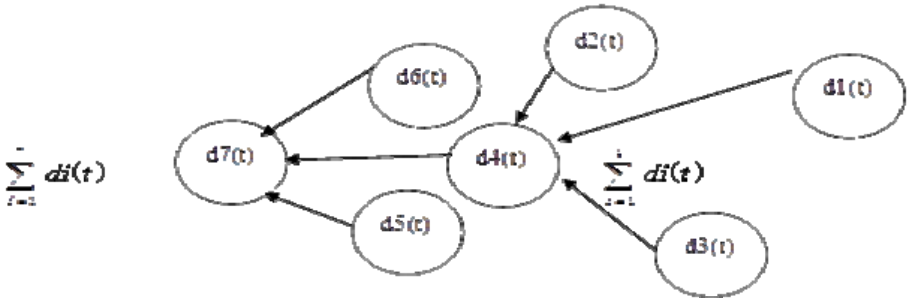

Fig. 5. The sum aggregation function

To realize privacy protection in WSN, the key is important to encryption as well as decryption. Different key generation mechanism and allocation mechanism have their own advantages and disadvantages. In this paper, the improved algorithm is a random key distribution strategy for hop by hop encryption. The steps of random key distribution strategy [20] are as follows:

Step 1: Generate a key pool, assuming that it contains K keys; nodes randomly gain k keys in turn.

Step 2: Each two neighbor nodes query each other to confirm whether they contain the same key between them, if they shared the same key, using it to encrypt data. The possibility, *P1*, that two nodes do not share a key is:

$$P_1 = ((K-k)!)^2 / ((K-2k)!K!) \quad \textbf{(2)}$$

So the possibility of two nodes connected $P_{connect}$=1-P_1

Step 3: If the two nodes do not have the same shared secret key, they can connect in more hops. Therefore, if attackers want to steal or tamper with the information, they must use the same key distribution strategy, if he takes out the key in k keys, he can use the key to decrypt the encrypted information passed between two nodes and the link is no longer safe. If an attacker gets the key, the possibility $P_{overhear}=k/K$.

Assume that there are totally 10000 keys, each node randomly selected 100 key, that is to say, k = 100, then the probability that eavesdropper can get the keys is 0.1%.

4.2 The description of the improved PSPDA algorithm

4.2.1 The improvement of the SMART

Combined with the ESPART algorithm which has a low energy consumption [21] and HEEPP algorithm based on difference divided [22] and the EEHA algorithm[23] , from the data traffic, privacy protection, data aggregation performance, integrity verification and other several aspects, we make corresponding improvements on SMART algorithm. The following is the introduction to PSPDA algorithm.

The different operation between different types of nodes: It refers to two aspects and one is that the aggregation nodes do not need to distribute data slice to neighbor nodes, reducing the communication overhead. The other refers to the structure features and using of characteristics of data aggregation tree nodes, the nodes in the aggregation tree whose indegree and outdegree are greater than or equal to 4 don't have to slice, only has to receive data. This reduces the amount of communication data in WSN. According to data aggregation tree, assuming that a node whose degrees are 4, even captured, you needs to break the rest of the four links to know the complete data, the probability $P_2=P_{overhear}^4$, which is a very small value. So we make the nodes whose degrees are greater than or equal to 4 do not need to slice in the improved algorithm.

Probabilistic segmentation is brought in the new algorithm when slicing: Each node slices the data according to a uniform probability distribution, the pieces of slices differs from each other as well as the exchange slice to neighbor nodes. For example, the maximum number J=4, in addition to the particular nodes mentioned in part 1, other nodes slice their data into two, three and four pieces in 33.3% probability respectively. The same way when J=5, other nodes slice their data into two, three, four and five pieces in 25% probability respectively. This approach can avoid excessive traffic, and at the same time increase privacy protection, because different nodes slice different pieces, even if some nodes are captured, it is hard to get all the data.

Add the data coefficient, the positive and negative factors: if we want to reduce the loss caused by collision as little as possible, we can set a coefficient to limit the data, making its distribution within a certain range. When the collision rate does not change, in this way we can effectively improve the accuracy of the aggregation. The designed data collusion coefficient F = (average of data)/J, each slice can be calculated like this: $slice_i = F \times r, r \in (0,1)$, i stands for the number i slice. Positive and negative factors can divide the sliced data into positive and negative according to the equal probability. We can assume that there are sliced data missing in the transmission, if missing data is negative, making the data aggregation result larger; and vice versa. . Considering that the attribute of the missing data are roughly at equal probability distribution, so it can have a compensation effect. If there are no data missing, the final aggregation result should be near to the theoretical value. Supposing that there are n nodes in network, each node will slice the data into J pieces and randomly send J - *1* slices to adjacent nodes , $Slice_1$, $Slice_2$,... , $Slice_{J-1}$, , the original information for each nodes is di (i = 1, 2... n). It can be summarized as: $Slice\ n = di \times (-1)^{(n-1)} \times r^n$, r $\in (0,1)$. And positive and negative slice factor can play a compensation effect. Combining the two aspects, we can calculate each slice for each node:

$$Slice_n = F \times r \times (-1)^{n-1} \times r^n, r \in (0,1) \quad \textbf{(3)}$$

In addition we can also consider making exchange time random in the algorithm, avoiding a clash when exchanging slices. In SMART algorithm, the time for exchanging slices for each node is the same as J. In the improvement program, we can change the time of each node to exchange information according to different amount and size of slices. We can effectively reduce potential conflict between neighbor

nodes. Under the assumption that each node's subdivision number is J, and the time window for sending slices is t_s, t_s is evenly divided into $J - 1$ pieces, $t_1, t_2 \ldots\ldots t_{J-1}, t_s$, and therefore the length of each paragraph is $(t_s - t_1)/(J - 1)$, for each node, the n-$th\ slice_n$, $t_n < t < t_n + 1$, we define t:

$$t = t_n + t_s \times \left| slice_n / d_i \right| / J \tag{4}$$

The delivery time when $J = 4$ is shown in the Fig. 6.

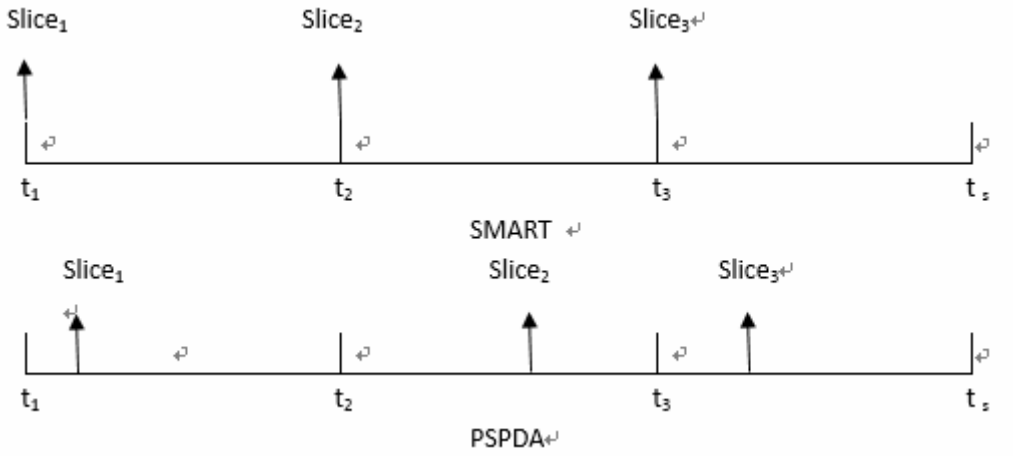

Fig. 6. The time for sending slices

We can also add data query mechanism which can reduce the possibility of missing information, increase the precision of aggregation. Data query mechanism can be set for each level with a waiting time t, the waiting time can be different from the levels. The higher the level is, the longer the waiting time. If longer than the waiting time, the node will send a data query message. If the intermediate node wait again and has not received the data from the fusion of child nodes after a period of time, it will give up the data of leaf nodes and send the received data to its parent node.

4.2.2 The procedures of the algorithm

The procedures of the algorithm are as follows:

Preprocessing stage: First of all we need to build a WSN, and determine the factors such as the size of simulation area, number of nodes, the communication radius, serial number, location of the base station and so on. At the same time, it is necessary to generate the corresponding matrix of child nodes and parent's nodes as well as the matrix recording the distance between nodes to distinguish the different types of nodes, to prepare for the construction of data fusion tree.

The parameters are set as follows: first, we set up the size of the simulation area, L, is 100 x 100, $NodeNums$ = 100, the maximum communication distance R is 30, coordinates for BS is (50, 50). Then the 100 nodes will be randomly arranged in this area with their respective id. The matrix of child and parent are 1 x 100 matrixes whose initial value is zero. And a 100 x 100 matrix recording distance should also be established.

When the distance between nodes and base station is less than communication radius R, the node type is "A" which represents a aggregation node, and we should link it with the base station and put 1 into the corresponding parent node matrix; Otherwise the node type is "N" which means common nodes, and the parent node matrix should be 0.

Finally build a 100 x 100 matrix to record the adjacent nodes for each node. Calculate the distance between any two nodes; if the distance is less than R, we put 1 into the corresponding location in the matrix, otherwise inf. This completes the PSPDA preprocessing stage.

The next stage is to build a data aggregation tree. All nodes in SMART algorithm share the same status, whereas in the improved algorithm different nodes have different operations. So the first thing is to draw the structure of data aggregation tree. During the deployment process, we distinguish the nodes as base stations, aggregation nodes and leaf nodes.

Building the data aggregation tree: First the base station in the center invited the nodes around, if the

node can communicate with the BS, it is made into an aggregation node with an identity "A", and others are "N". If the node has no parent node, the BS will become its parent node and link to others. Then 1 is added into the parent matrix to indicate the existing of the parent node.

Then we start from each aggregation node, if the value in the child matrix of the node is less than 5, send a request to nodes around. If the value in the parent matrix of the received nodes is 0, connect the two nodes and, at the same time establish a 100 x 100 matrix recording the connections between nodes. If there is a connection between *node i* and *node j*, the value in *link (i, j)* should be 1, otherwise 0. The corresponding value in child matrix plus 1, and the requested node's parent matrix should update in time.

This applies to the rest of nodes by looking for the nodes with no parent node, then deal with them in the same way.

Before simulation specific parameter settings and their corresponding definition should be defined (see Table 1).

Table 1. The specific parameter

the parameter in simulation	The meanings of the parameters
L	The size of the WSN simulation area
R	The biggest communication radius of node
BSx,BSy	The location of the base station
Parent	1×100 matrix to record if node has a parent node
Chil	1×100 matrix to record if node has a child node
dist	Calculate the distance between the nodes
matrix	100×100 matrix to calculate if two nodes are neighbors
link	100×100 matrix to record whether two nodes are linked

For calculation, data collected in each node are randomly assigned from 1 to 100, and displayed in the figure.

Next we compare the communication overhead: calculate the number of slices of each node to determine the outdegree and establish a 6 x 100 matrix to record them when J is from 3 to 8. Categorize all nodes and calculate the indegree and outdegree of each node in the aggregation tree and deposited in the corresponding matrix. In the improved algorithm, it will not slice if the sum of indegree and outdegree is more than 4 and we should set the number of slices of leaf nodes accurately, making it follow uniform distribution. For example, if J=3, the nodes will divide data into 2 or 3 pieces with both equal 50% probability in addition to the aggregation nodes. Assuming that J=5, each nodes will divide data into 2, 3, 4 or 5 pieces with 25% probability. And then calculate the ideal communication overhead of the two algorithms by the outdegree matrix, *C1* stands for the communication overhead of the SMART algorithm, while *C2* stands for the improved algorithm. Compare *C1* with *C2* when J is different.

Then we compare the privacy protection degree between two kinds of algorithm: first set J = 3, we know that, in SMART algorithm, each node divides the data into three slices and the outdegree is two. Calculate the size of slices of each node separately, and put the values into the first two lines of a 5 x 100 matrix. Then calculate the *indegree1* which represents the indegree of SMART algorithm and *indegree2* standing for the improved algorithm. Calculate the probability P_1 and P_2 when the indegree is k in each scheme. Assuming that the probability that link is cracked is q, calculate and compare the privacy protection degree.

The last stage is the comparison of the data aggregation precision: assuming that the missing probability because of sending the slice at the same time is p, validate the data aggregation precision of two algorithm along with the change of p. Table 2 shows the pseudocode of the PSPDA algorithm when J = 3.

5 Simulation and Performance Analysis

We simulate the improved algorithm to verify the improvements compared with the SMART algorithm in terms of communication consumption, the degree of privacy-preserving and the aggregation precision.

Table 2. The pseudocode of the PSPDA algorithm when $J = 3$

```
1. Preprocessing
2.   NodeNums=100,R=30,L=100
3.   Construction of the tables chil and parent
4.   Construction of the table matrix and link
5.   Computation of dist
6. If dist<R
7.         Then  Node(i).type='A',parent(1,i)=1
8.     Else Node(i).type='N', parent(1,i)=0
9. End of if
10.   Construction of the tree of data aggregation
11.   For i=1 to 100
12.       For j=1 to 100
13.       Search the nodes without a parent node in a jump range from the
   aggregational node
14.           If  (Node(j).type=='N')AND (chil(1,i)<5) AND (parent(1,j)==0)
15.               link(i,j) = 1, chil(1,i)=chil(1,i)+1, parent(1,j)=1
16.           Else link(i, j) = 0
17.           End of if
18.       End of for
19.   End of for
20.   Comparison of data communication
21.   Comparison of privacy protection
22.   Comparison of accuracy of data aggregation
```

5.1 Simulation of communication consumption

The energy consumption in WSN is mainly divided into communication consumption, accounting for most of the energy consumption, and calculation consumption, which has a close relationship with the performance and existing time of the WSN. In the matlab simulation, we choose the number of the packets under the ideal state as an indicator to measure the consumption of communication as shown in Fig. 7 below:

Red star with dashed lines in the Fig. 7 stands for the SMART algorithm, while the blue square with dashed lines represents improved algorithm.

We can analysis that in the SMART scheme, every node have an operation of slicing thus *C1 = (J - 1)* × *NodeNums*, while the improved algorithm is the sum of *outdegree* of all nodes, namely the sum of the first row of the *outdegree* matrix. Take *J*=3 for an example, *C1*=100×(2-1)=200, *C2* equals the sum value of the first row of the *outdegree* matrix. We can see from the Fig. that the bigger the *J* is, the greater the difference of their energy consumption is. So the proposed improved algorithm is significantly lower than the SMART algorithm in terms of the communication consumption.

5.2 Simulation of privacy protection degree

To measure the WSN privacy protection performance, we choose the privacy protection degree. The way to invade the privacy of a node in the network can be divided into eavesdropping and conspiracy, and privacy protection degree represents the probability of the original data obtained by compromised neighbor nodes. In the simulation, in order to facilitate calculation, assuming eavesdropping and conspiracy probability is equal to q

Define privacy protection degree Q as follows:

$$Q = q^{J-1} \sum_{k=0}^{d\max} p(indegree = k) \times q^k \tag{5}$$

Among them, $p(in\deg ree = k)$ is the possibility that the indegree of the node is k, d_{max} is the max indegree of the node. For the improved algorithm, the slice number are not fixed, so privacy protection degree can be calculated by sum of the degrees of the nodes to present the probability of privacy information missed. For convenience, this paper only simulates privacy protection degree of the two algorithm when *J*=3, 4 and 5.

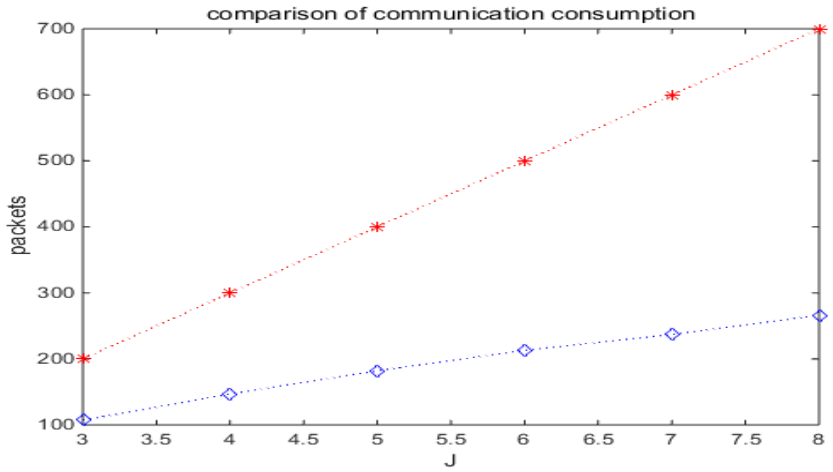

Fig. 7. The comparison of communication consumption when *J* is different

As is shown in Fig. 8 to 10, the possibility that the link between the nodes is cracked is 10%:

When *J*=3, the probability that the privacy information are exposed in SMART algorithm is $1.8\times10^{-3,}$ whereas the improved algorithm is 1.4×10^{-3} .

When *J*=4, the probability that the privacy information are exposed in SMART algorithm is $8.3\times10^{-5,}$ whereas the improved algorithm is 7.1×10^{-3} .

When *J*=5, the probability that the privacy information are exposed in SMART algorithm is $2.5\times10^{-6,}$ whereas the improved algorithm is 2.8×10^{-6} .

From this we can see the improvement algorithm performance of privacy protection is better than the SMART algorithm.

We can analyze from the figures with the maximum number of slicing *J* increasing, the leaked probability of two kinds of algorithm is smaller and smaller, but the privacy protection degree of improved algorithm is slightly higher compared with SMART, and with *J* increasing, the degree of privacy protection gap between them is smaller and smaller and the communication overhead is increasing too. At the same time, because of the different slice number, it's hard to restore all the original data for a captured node. This shows that the privacy protection performance of the improved algorithm in the network is better than SMART algorithm, especially when *J* is bigger.

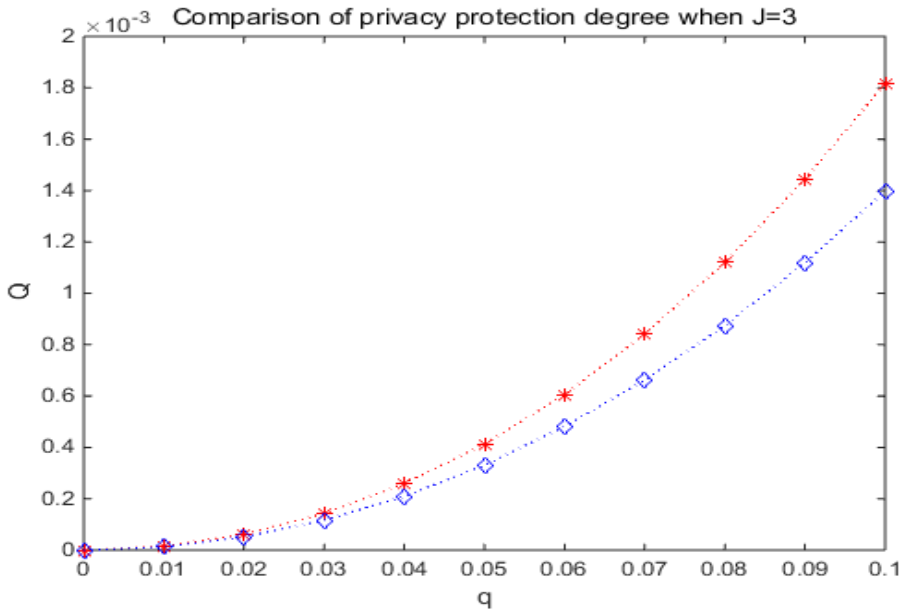

Fig. 8. *J*=3

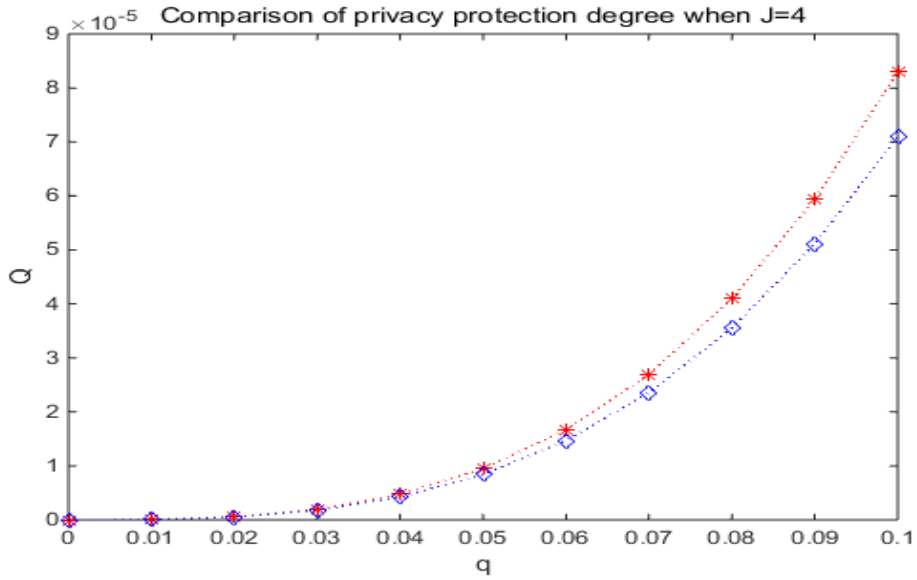

Fig. 9. *J*=4

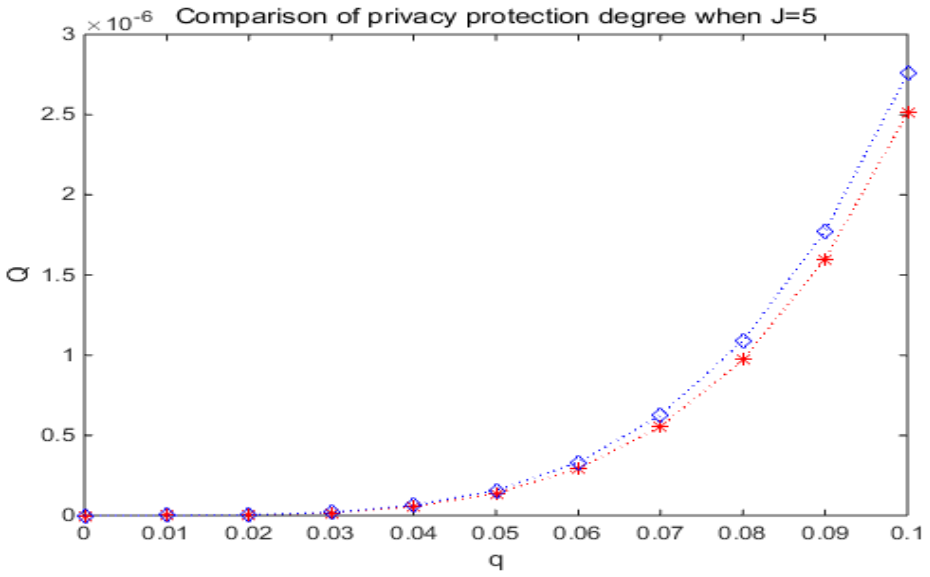

Fig. 10. *J*=5

5.3 Simulation of the precision of data aggregation

The accuracy of data aggregation or the rate of the receiving slices is one of the important indicators in data fusion performance in WSN. Ideally, in the process of transmitting data packets, if there is no collision or lost, the data aggregation accuracy should be 100%. The accuracy of data aggregation is defined as the rate of the sum of the data after aggregation and the sum of the original data. When *J* = three, four or five, the aggregation accuracy are shown in Fig. 11 to Fig. 13.

As we can see from the Fig. 11 to Fig. 13, when the data collision loss probability is 0, both aggregations accuracy are 1, but with the increasing of the probability of data collision, the accuracy of the data aggregation is reduced. In general, as the probability of collision loss up, the difference of performance between two algorithms is more obvious.

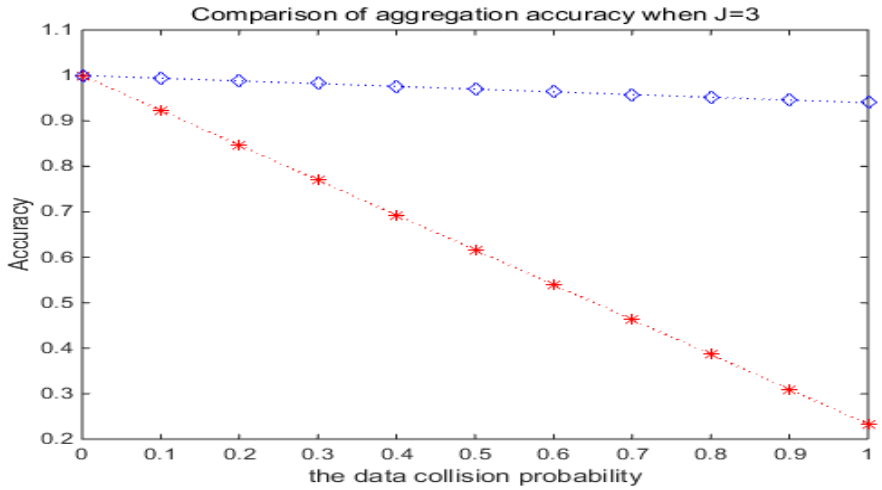

Fig. 11. The aggregation accuracy when *J*=3

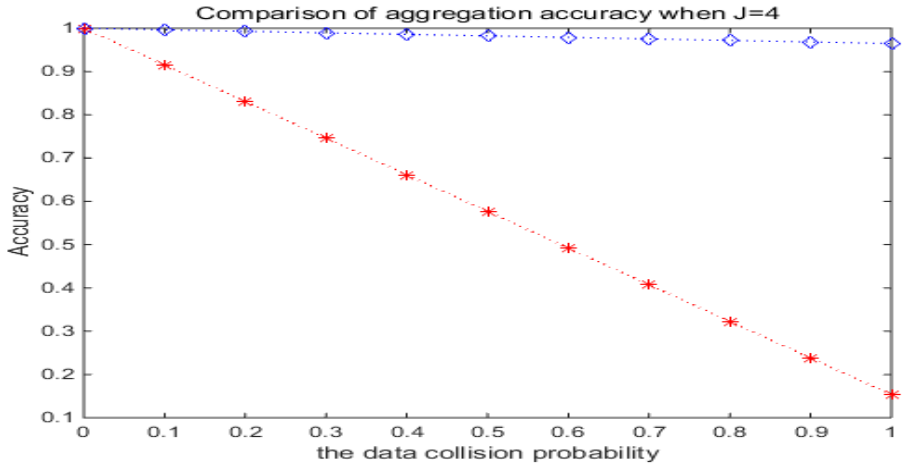

Fig. 12. The aggregation accuracy when *J*=4

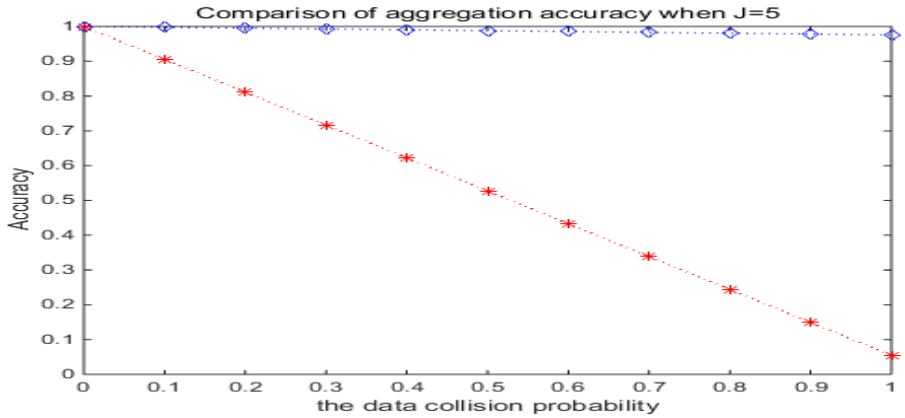

Fig. 13. The aggregation accuracy when *J*=5

We can see that the improved algorithm is optimized on the slice number and size, join the collusion data coefficient at the same time to reduce the damage caused by the collision, introducing the random time factor in the improved algorithm to send the slice in different time. Compared with the SMART algorithm, the data collision probability of improved algorithm is far below the SMART algorithm and what we show in the Fig. 13 is under the assumption of the same collision probability, so the proposed algorithm has a better performance than SMART.

5.4 Analysis of the performance

From the simulation results we can see that the data exchanged between nodes has a sharp reduction compared with SMART, making the data traffic and the probability of collision reduce. At the same time, each node follows the uniform distribution make the attacker more difficult to get all the data, thus it is difficult to determine whether to restore the original data of a node. So the privacy protection performance for the whole WSN gets significantly enhanced.

In Fig. 7 it shows that the bigger J is, the greater the exchange of information, the improved algorithm can greatly reduce the energy consumption of WSN network.

Compare with the Fig. 8-10 we can see that with the increase of the link-cracked probability, the greater the chance that privacy of data are cracked For example, when the link-cracked probability p = 0.1, J = 3, the data-leaked probability in SMART is 0.9×10^{-3}; When J = 4, the probability is 6×10^{-5}. However, we need to balance the relationship between the energy consumption and the data privacy protection.

What Fig. 11-13 show us is that the proposed algorithm made up for the loss caused by collision and aggregation accuracy is much higher than SMART algorithm.

Because we cannot use MATLAB to carry out all necessary simulation and tests in WSN, there are limitations in this evaluation. However, what we have observed from the work carried out so far demonstrates that the improved algorithm has obvious advantages compared with the SMART algorithm. It has proved that PSPDA algorithm is suitable for WSN and has lower energy consumption, higher privacy protection and accuracy of the data aggregation result.

6 Future Work

There are still many limitations to be overcome in privacy-preserving technology in data aggregation. Some of the suggested ideas are:

1. The privacy protections in future may involve combination of multiple encryption methods, depending only on a single way is not reasonable. For example in cryptography, using hybrid Encryption may have a surprising result. What's more, with the coming of the Internet of things and the era of big data, data query and data anonymous technology could be widely used.
2. Because the present aggregation type supported is very limited, a new privacy protection technology in the future should be able to support a variety of data aggregation type.
3. Because WSN will be applied in all fields in the future, there is a urgent need to develop privacy protection technology to meet the needs of the new applications.
4. There are still many factors such as communication performance, privacy protection performance, the accuracy of the data aggregation results, as well as integrity verification to be research in order to improve privacy protection algorithms for WSN.

Acknowledgement

This research is supported by Fundamental Research Funds for the Central Universities (2015YJS027), National Natural Science Foundation under Grant 61371071

References

[1] D. Culler, D. Estrin, M. Srivastava, Overview of Sensor Networks. IEEE Computer 37(8) (2004) 41-49.

[2] R. Bista, J.W. Chang, Privacy-preserving data aggregation protocols for wireless sensor networks: a survey, Sensors 10(5) (2010) 4577-4601.

[3] L. Cui, H. Ju, Y. Miao, T. Li, W. Liu, Z. Zhao, Overviews of wireless sensor networks, Journal of Computer Research and Development 42(1)(2005) 163-164.

[4] G. Tan, Research on source node location privacy protection in wireless sensor networks, [dissertation] Hefei: Anhui Uni-

versity, 2014.

[5] A. Ukil, Security and privacy in wireless sensor networks, in: Y.K. Tan (Ed.), In Tech Croatia, 2010, pp. 395-418.

[6] F. Yang, Research on secure multi-party computation, [dissertation] Jinan: Shandong University, 2007.

[7] Y.-J. Fan, H. Chen, X.-Y. Zhang, Data privacy preservation in wireless sensor networks, Chinese Journal of Computers 35(6)(2012) 1132-1134.

[8] J. Xu, G. Yang, Z. Chen, Q. Wang, A survey on the privacy-preserving data aggregation in wireless sensor networks, China Communications 12(5)(2015) 162-180.

[9] W. He, X. Liu, H. Nguyen, K. Nahrstedt, T. Abdelzaher, PDA: privacy-preserving data aggregation in wireless sensor networks, in: Proc. of 26th IEEE International Conference on Computer Communications, 2007.

[10] P. Qian, M. Wu, A privacy preserving method in WSN, 〈 http://www.cnki.net/KCMS/detail/detail.aspx?QueryID=0&CurRec=1&filename=DXKX201301007&dbname=CJFD2013&dbcode=CJFQ&pr=&urlid=&yx=&v=MTAzMzZZNFI4ZVgxTHV4WVM3RGgxVDNxVHJXTTFGckNVUkx5ZVplVnZGeWpuVmIzQklUWEFkckc0SDlMTXJvOUY=〉, 2013.

[11] W. He, H. Nguyen, X. Liu, K. Nahrstedt, T. Abdelzaher, iPDA: an integrity—protecting private data aggregation scheme for wireless sensor networks, in: Proc. of IEEE Military Communication Conference (MILCOM 2008), 2008.

[12] Y.-Jian Fan, H. Chen, X.-Y. Zhang, Data privacy preservation in wireless sensor networks, Chinese Journal of Computers 35(6)(2012) 1135-1137.

[13] J. Xu, G. Yang, Z.-Y. Chen, H.-Y. Wang, G. Yang, Research of privacy-preserving technology in wireless sensor network data aggregation, Computer Engineering 38(15)(2012) 134-136.

[14] Y.-J., Fan, H. Chen, X. Zhang, Data privacy preservation in wireless sensor networks, Chinese Journal of Computers 35(6)(2012) 1135-1137.

[15] Y. Chen, C. Fu , J. Xu, G. Yang, Lightweight privacy-preserving data aggregation algorithm, Journal of Computer Applications 34(8)(2014) 2336-2337.

[16] Y.-J. Fan, H. Chen, X.-Y. Zhang, Data privacy preservation in wireless sensor networks, Chinese Journal of Computers 35(6)(2012) 1137.

[17] J. Li, J. Cui, Elliptic curve encryption algorithm and the application, China Academic Journal Electronic Publishing House (11)(2014) 55-57.

[18] J, Xu, G. Yang, Z.-Y. Chen, H.-Y. Wang, G. Yang, Research of privacy-preserving technology in wireless sensor network data aggregation, Computer Engineering 38(15) (2012) 136-137.

[19] C. Liu, Research on Secure Data Aggregation in Wireless Sensor Networks, Beijing Jiaotong University, Beijing, 2013.

[20] G. Yang, A.-Q. Wang, Z.-Y. Chen, J. Xu, H.-Y. Wang, An energy-saving privacy-preserving data aggregation algorithm, Chinese Journal of Computers 34(5)(2011) 792-800.

[21] A.-Q. Wang, Research on data aggregation algorithm in wireless sensor network, Nanjing University of Posts and Tele-Communications (2012) 31-38.

[22] C.-X. Liu, Research on secure data aggregation in wireless sensor networks, Beijing Jiaotong University 2013, 59-65.

[23] H.J. Li, K. Lin, K.Q. Li, Energy-efficient and high-accuracy secure data aggregation in wireless sensor networks, Computer Communications 34(4)(2011) 591-597.

[24] N. Li, N. Zhang, S.K. Das, B. Thuraisingham, Privacy preservation in wireless sensor networks: a state-of-the-art survey, Ad Hoc Networks, 7(8)(2009) 1501-1514.