Additive aggregate function-based data privacy protection algorithm

Yan-Sheng Chen*
School of Eco-environmental Technology,
Guangdong Industry Polytechnic,
Guangzhou, China
Email: 362070614@qq.com
*Corresponding author

Tzong-Ru Lee
Department of Marketing,
National Chung-Hsing University, Taiwan
Email: trlee@dragon.nchu.edu.tw

Abstract: Data privacy protection is a key problem in the research and application of the internet of things that in wireless sensor networks has broad application prospects, such as agriculture environmental monitoring, healthcare, etc. For the core part of the internet of things, wireless sensor network, it is a very challenging task to provide effective privacy protection in the process of data transmission. In this paper, a new type of data privacy protection strategy based on the additive aggregation function is proposed (hereafter referred to as DPPA). On the one hand, applying the superposition property of the node data, the paper introduces the class aggregation protocol. On the other hand, taking advantage of the algebraic properties of the polynomials and the correlation properties of the superposed data, it can reduce the data of the header files of the data packages, and reduce the data size, while improving the privacy preservation. Simulation results show that DPPA can effectively protect data privacy, and get accurate data fusion results, while reducing the amount of data traffic.

Keywords: internet of things; wireless sensor network; WSN; privacy preserving; data aggregation; superimposed correlation.


Biographical notes: Yan-Sheng Chen is an Associate Professor at the Guangdong Industry Technical College, working in the field of computer applications/data mining and machine learning. He is also teaching courses in network technology, the principle of network communication, cyber source application and security technology, computer hardware maintenance and repair, computer science and computer application.
1 Introduction

Wireless sensor network (hereinafter referred to as WSN) is composed of a large number of distributed sensors and it is used in the perception of valuable physical information in the daily life of people. It converts the physical information into data that is suitable for network transmission and brings the feedback onto server data centre and realises the real-time monitoring and data monitoring on target area after decoding of background data. However, the important information resources in WSN like location information and physical data information perceived by the sensor nodes are sensitive data needing special protection, such as the life information of body temperature, blood pressure and heart rate of patients in the elderly health management system. Therefore, The scholars such Shao (2017) and Fei et al. (2017) believe that to carry out effective privacy protection of the sensitive information perceived by sensors is one of the primary targets of the research on WSN technology and is also one of the core problems needing to be solved in the practical application of WSN.

In WSN, the sensors of nodes can produce large amount of sensory information with time and actual situation of physical quantities of sensing field. And the quantity of transmission data can directly affect the energy consumption of nodes. Studies have shown that, when adopting Micadot node communication, the energy consumption of sending 1 bit of data will be about 4,000 nJ and the energy consumption for processor to execute one instruction will be about 5 nJ (Oreku, and Pazynyuk, 2016). Therefore, reducing the data traffic of nodes and transmitting valuable data as well as preliminarily processing data rather than direct transmission of original data are more meaningful to the power consumption control of the entire sensor network. Normally, if repeated information is sent to the same node or the adjacent nodes receive repeated data, it might lead to valueless energy consumption of node sensors. So the effective fusion of node data becomes particularly important. In recent years, many researchers have spent a lot of energy and have made a great number of achievements on data fusion of WSN.

Sensor network technology is widely applied in military and civil field, even in all kinds of privacy scene monitoring of daily life, so that privacy data protection work gets more and more attention. Usually, data fusion process will not encrypt privacy data and the sensors of adjacent nodes can mutually obtain any data from each other. Therefore, on the basis of data fusion, the privacy protection on sensitive data is a key step for the security of WSN. In order to protect the security of node data, some data privacy protection mechanisms have been proposed, such as ESPDA (Fissaouii et al., 2018) and SRDA (Gilbert et al., 2018). But these algorithms will limit the data types in the process.
of data fusion or will cause additional communication overhead and increase communication traffic. In addition, The scholars such Yavuz (2016), Baldi et al. (2016), Zhang et al. (2016) and Apostolaki et al. (2017) believe that attackers can still parse privacy data after intercepting the encrypted files and the privacy protection effect is not ideal.

In order to solve the problem of privacy data protection, in general, solutions can be sought from the following several aspects:

1. In the process of communication, implement encryption processing of privacy data. The privacy protection scheme to be designed should be able to realise the transparent transmission of the other nodes and links by target node data when the attacker intercepts some nodes in WSN and even monitors the whole network (Chen et al., 2015; Sethi et al., 2017).

2. Embark from the algebraic property of data after the fusion computation of node data and be capable of correcting the data within error margin (Zhang et al., 2017).

3. While improving the degree of security protection of privacy data, optimise communication, reduce the operational energy consumption of network, especially the sensor nodes with limited energy and further enhancing the survivability of WSN (Cheng and Liang, 2016).

On the basis of the above-mentioned design ideas, the privacy protection strategies put forward by many scholars generally complete end-to-end encryption processing directly and perform encryption operation inside WSN. The adoption of this privacy homomorphic encryption method based on additive polynomial can let family and header files perform additive operation on encrypted data. Since then, some homomorphic encryption-based data fusion mechanisms can achieve a higher level of security protection.

This paper presents a kind of dynamic and extensible privacy data protection algorithm that is based on additive aggregate function. It builds the connection between the privacy data in WSN and the data after fusion through algebraic property and also utilises the adjacent node data to encrypt target node data. When any node, any transmission link or part of sensor network is attacked, it can effectively protect data privacy. On the other hand, it can effectively recover data damage too. The sensor network can accurately capture the data after fusion and ensure that the privacy node data will not be leaked to other nodes. Compared with other data privacy protection algorithms, this algorithm has the characteristics of low computational complexity and small communication traffic. Moreover, its privacy protection is also strong. Its safety performance and fusion efficiency have been improved on the basis of the existing algorithm. At the same level of privacy protection, its data communication (traffic) is reduced by over 40% than that of traditional AMART algorithm.

The first part of this paper discusses the related research work in the current privacy data protection and analyses the protection algorithms under two different cases: the respective privacy protection strategy and performance of PASKOS algorithm and PASKIS algorithm; the second part describes the mathematical model of data fusion protection algorithm designed by this paper; the third part carries out simulation and analysis on DPPA, compares its performances in privacy protection, work efficiency and data damage resumption with those of other algorithms and analyses its optimisation effect.
2 Related work

In WSN, the energy supply of sensor nodes is usually limited. In order to save energy of nodes, the initial data must process the communication (traffic) overload in network after fusion (Simonsen et al., 2008). In terms of data fusion processing, a variety of processing mechanisms and running algorithms have been put forward. In the designs of these algorithms, the premise is that all sensor nodes and their communication are safe. However, in practical application, wireless sensors are exposed to multiple sorts of complex environments. Attackers might intercept part of sensor nodes, transmission links and even part of sensor network bee and then get the privacy data via parsing the key.

PASKOS algorithm can protect the target node data through the nearby sensor node data and has good privacy protection performance. In the event that attackers intercept some nodes in the network (non-communication link), it can effectively protect node privacy data. In contrast, PASKIS algorithm can realise the strategy of protecting privacy data in the case that communication link and some nodes are intercepted at the same time. SMART algorithm (Lezama et al., 2014; Ren et al., 2017) adopts hop-to-hop data fusion method and point-to-point encryption processing to protect privacy data. All the above-mentioned algorithms do fragmented processing on target node privacy data and then get them mutually overlapped and fused with each other. Their computing complexity is higher and their data communication (traffic) is larger. Moreover, the fault-tolerant capability of SMART algorithm is poorer. Once data is missing or data damage reaches a certain extent, irreparable consequences will be caused, which can seriously affect the precision performance of sensor network.

In terms of privacy protection, the peer-to-peer concept and proposed two kinds of mainstream protection mechanisms: one is based on data crosstalk technology. It regularly inserts a group of random sequence into privacy dataflow and uses the distributive features of random sequence at the receiving end to restore original data. Meanwhile, it employs random data to conceal real data and to further achieve the purpose of protecting privacy. Nevertheless, data crosstalk technology has certain defects and it fails to produce accurate fused data due to the randomness of inserting data. For this, Kargupta et al. put forward that certain sequence can be applied for data crosstalk. Research results showed that data protection effect was not so ideal (Oliveira et al., 2011).

The other privacy protection mechanism is data mining technology based on SMC (safety). It makes the problem of data privacy protection equivalent to the computational processing of multiple input data streams. Usually, it issues a public key for data encryption and the processor uses this key to carry out the operations of encoding and decoding so as to attain the aim of protecting data. This method always has a heavy computation and is unable to adapt to WSN that has limited energy.

The privacy data protection algorithm proposed by this paper is improved on the basis of SMART algorithm and it has the following advantages:

1. data privacy protection effect can meet the requirement that it is only visible to the legitimate users of WSN
2. within the scope of certain damage, data fusion have error correcting capability and can transmit data accurately
3. it has high efficiency, strong stability and strong adaptability.
3 System model

3.1 Sensor network and data fusion model

In this paper, WSN is expressed with the connected graph $P(\rho, C)$, the sensor node in network is $\rho$, the connection link between nodes is $C$ and the number of sensor nodes in WSN is $N$. Data fusion function is defined as follows:

$$y(t) = P(d_1(t), d_2(t), \cdots, d_N(t))$$  \hspace{1cm} (1)

Here, $d_i(t)$ represents the data collected by the $i^{th}$ sensor node in time $t$, $P$ represents the fusion function and the output of fusion function is data flow. Considering the actual situation of the application of WSN, data fusion technology should pay great attention to the problem of energy consumption of system and the fused data flow should meet the requirements like reducing data size as much as possible, restricting the bandwidth of wireless link and small consumption rate of transmission power of data packet.

The data fusion function adopted by this paper is as follows:

$$Pd_i(t) = \sum d_i(t) + vd_i(t)$$  \hspace{1cm} (2)

Here, the operator $\sum d_i(t)$ is an additive operation of node data and its function is to implement data aggregation. The operator $vd_i(t)$ is a random sequence formed by a data stream and it obeys normal distribution. Its means is 0 and its variance can be set. It can realise the function of dynamic configuration and improve the self-adaptability of protection strategy.

3.2 Data encryption

In general, for data encryption process, firstly the setting of keys must be conducted. The digital features of keys will directly affect the safety, effectiveness and work efficiency of the system. Therefore, it is critical to reasonably configure the keys that are applicable to different security levels.

In this algorithm, the setting of keys shall comply with the following three principles:

1. Key distribution: first of all, one key pool $K$ needs to be equipped. To say it informally, it means that a series of data strings shall be formed through a certain mathematical operation, their corresponding algebraic properties shall be characterised and each sensor facing the network shall be assigned with the designated key. The distribution way is always randomly selected and the number of keys for each node to get is $k$.

2. Key recovery: each sensor node can sense the key of adjacent node and they exchange differential data after key validation and decoding. At this point, if two adjacent nodes share one key, it indicates that the link connecting two nodes is safe and reliable.

3. Estimation of link key: link key is defined as follows: two adjacent sensor nodes that share non-public key can be interlinked by two or more than two reliable links.
This pair of nodes of non-public key can use link key recovery algorithm to estimate a public key. After a pair of nodes gaining the public key, the probability for any other node to use this key. As for a large number of sensor nodes, this probability can meet the demand of protecting privacy for a large number of sensor nodes.

### 3.3 Data fusion

This part introduces privacy data fusion strategy based on additive property. It is mainly divided into three steps: data aggregation, eigenvalue calculation and the aggregation of fusion results. The letters represent the sensor nodes distributed in the network. Before data transfer, each node has the default of being cluster-head node itself and releases the information to adjacent nodes. After a certain time interval, the next level of nodes receive the information delivered by adjacent nodes and joins in this cluster by means of giving feedback signal J to sending nodes. It is the result after grouping. Here the cluster-head nodes of each cluster are respectively \(i\). Then, implement data fusion in the unit of cluster. Among them, each node is done with the following operations:

\[
\rho_i = \alpha + \beta i + \gamma i^2
\]

\(\alpha\) is the privacy data of node X, \(\beta\) and \(\gamma\) are keys and node \(i\) is cluster-head node.

### 3.4 Privacy protection algorithm

In SMART privacy protection algorithm, nodes cut privacy data into data fragments and then send them to different collectors after encoding in the unit of piece of data. The collectors will classify the received data and send them to the stack. When the stack receives data, it will compute and judge if it is the final fusion result. This method makes improvement on SMART algorithm and attains the dynamic configuration of encryption process. Keys can be flexibly selected according to the security level and the work efficiency of WSN can be further improved.

**Step 1 Initialization**

Suppose the sensor node in WSN is \(i(i = 1, \ldots, N)\), randomly select a group of nodes \(S_i\) and the cluster-head node privacy data inside the cluster is \(d_i\). Here, \(J\) represents encryption order and \(J + 1\) is the order of corresponding key polynomial. Make \(m_{ij}\) represent the raw data sent from node \(i\) to node \(j\). When there does not exist data transmission, \(m = 0\). The data fusion result caused by cluster-head nodes can be expressed as follows:

\[
M = \sum_{i,j} m_{ij}
\]  

**Step 2 Data encryption**

After one node receives the fusion data delivered by another cluster-head node, the system will wait for a preset time period \(T\). Then the system will receive the enabling information \(S\) sent by the node after \(T\), process the currently received data according to the enabling information, and also fuse the information.
delivered by this cluster node to update the enabling information and then to send it. The encrypted data is as follows:

\[ A = \sum_i M + S \]  

Step 3  Data analysis and decoding

After the server receives the data sent by cluster-head node, it can decode the data according to the information of enabling signals and can finally get the data of each sensor node.

4  Performance analysis

This part mainly carries out analysis on aspects such as the data communication (traffic), safety performance, error correcting capability and energy consumption of DPPA. This paper adopts MATLAB software for analysis. The basic parameters of WSN are as follows: 200 nodes are randomly distributed, link noise is White Gaussian Noise, the environment is standard indoor environment, noise figure (NF) is 3.8 dB, the longest communication distance between nodes is 50 m, the ceiling of number of cluster nodes is 20 and the ceiling of key complexity is of nine orders.

4.1  Data communication

At the nodes for algorithm initialisation, the communication (traffic) during the process of formation of clusters is very fewer compared with data transmission volume after that. In order to simplify simulation process, at the initialisation phase of this algorithm, the data communication (traffic) can be negligible and it just needs to record the location information and grouping situation of each sensor cluster on the server.

In contrast with SMART algorithm, in the process of simulation, data communication (traffic) measures the total number of data packets to be sent within time. Here the changing rule of data communication (traffic) with the system running time is just as shown in Figure 1. Through the curves in the figure, it can be obviously seen that, compared with SMART algorithm, the data communication (traffic) of the algorithm put forward by this paper obviously declines under the same running time. From the aspect of algorithm principle, it is not difficult to analyse the advantages of this algorithm in data communication (traffic). For SMART algorithm, its protection mechanism is the encoding and decoding after fragmentation of privacy data. But this method is to conduct additive aggregate of privacy data and then finish encoding and decoding after compressing multiple node data through the extraction of eigenvalue. The compression process will reduce data communication (traffic).

4.2  Safety performance analysis

As for WSN, if any data packet of nodes is intercepted, whether the privacy data carried by it to be parsed out by the attacker or not will directly affect the safety performance of this network. The comparison between the safety performance of this algorithm and that of SMART algorithm is as shown in Figure 2. The advantages of the algorithm proposed
by this paper (curve a) in safety performance can be clearly seen from this figure. The safety performance of SMART algorithm (curve b) falls sharply with the increase of probability of being intercepted. However, this algorithm can effectively prevent data packets from being successfully parsed after interception. This depends on the enabling information in the process of encryption. As long as the enabling information is not parsed, even if the attacker has intercepted many data packets, the attacker cannot decipher them after grade checking and contrast.

Figure 1  Data communication (see online version for colours)

Figure 2  Safety performance (see online version for colours)
4.3 Error correcting capability

Under wireless communication environment, it is inevitable for data packets to have error codes in the process of transmission. This part will analyse the error-resistance performance of DPPA, namely the error correcting capability of the system. The curve for its error correcting capability is shown in Figure 3. It can be seen from the change rule of the curve in the figure that: with the increase of bit error rate of received data packets, the probability for data to be correctly parsed declines a bit. When bit error rate is less than $8 \times 10^{-3}$, the error correcting capability of the system is above 70%. After bit error rate gradually increases, the descending rate of the error correcting capability of the system will get accelerated. When bit error rate is over $5 \times 10^{-3}$, the error correcting capability of the system will drop below 50% and this will have an impact on the system’s correct parsing of privacy data carried by data packets.

Figure 3 Error correcting capability (see online version for colours)

5 Conclusions

This paper puts forward a novel kind of data fusion privacy protection algorithm in WSN and it is improved on the basis of SMART algorithm. Compared with SMART algorithm, it makes improvements in aspects like data communication (traffic), privacy data protection performance, the safety performance of system and error correcting capability of system. In future, further improvement will be made on this algorithm. And studies on the fusion of key and other data such as its system-level combination with wireless communication transmission will be conducted from the perspective of the corresponding constellation map of the real part and the imaginary part of key so as to solve the problem of big uncertainty introduced by the randomness of key distribution and to make this protection strategy more effectively combined with other data fusion methods.
Additive aggregate function-based data privacy protection algorithm

addition, great efforts will be made to resolve the remaining problems in ESPDA and SRDA.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (NSFC) (11471342)/Province natural science fund of Guangdong (2016A030313829)/Education department of Guangdong province in 2017 provincial key platform scientific research project (2017GWTSCX001)/Guangzhou City Philosophy and Social Science ‘13th Five-Year Plan’ 2018 Annual Fund Project (2018GZYSB103).

References


