Privacy Protection Scheme for Data Aggregation in Wireless Sensor Networks Based on PRDA+ Protocol

https://doi.org/10.3991/ijoe.v14i11.9503

Bohua Guo(✉), Yanwu Zhang
Shandong Binhai University, Shandong, China
bohuaguo2193@163.com

Abstract—To improve the data aggregation privacy protection scheme in wireless sensor network (WSN), a new scheme is put forward based on the privacy protection of polynomial regression and the privacy protection method based on the homomorphic encryption. The polynomial data aggregation (PRDA+) protocol is also proposed. In this scheme, the node and the base station will pre-deploy a secret key, and the random number generator encrypts the random number for the seed through the private key, which protects the privacy of the data. Then, by comparing the decrypted aggregate data through the correlation between the two metadata, the integrity protection of the data is realized. A weighted average aggregation scheme that can be verified is proposed. In view of the different importance of user information, the corresponding weights are set for each sensor node. EL Gamal digital signature is used to authenticate sensor nodes. The results show that the signature verification algorithm enables the scheme to resist data tampering and data denial, and to trace the source of erroneous data.

Keywords—data aggregation, privacy protection, WSN

1 Introduction

With the advent of the big data era, data acquisition, collection, collation, analysis and mining become more and more important. The wireless sensor network (WSN) has a variety of types of sensors, such as collecting temperature, location, and light intensity, and can collect data in different application environments. Then, it is necessary to process and analyze the data collected to get the information expected. WSN is a special network system composed of several inexpensive micro sensor nodes. Nodes are often deployed in the monitoring area through wireless communication. The purpose is to collaboratively perceive, collect and process the information of the perceived users in the covered area, and send the results to the analyst. The three elements of WSNs are sensors, users and data analysis operators. With the rapid development of the large-scale WSN technology, its application field is also more and more. Nowadays, WSN has played an important role in national defense military security, environmental detection and control, smart home and many other fields.

At present, the research on WSNs is mainly focused on routing algorithms, energy management, and node location, but the research on data aggregation technology in WSNs is not much. In WSNs, the energy of the sensor nodes is limited, the computing power is not strong, and the storage capacity is small. Therefore, it is not practical to use a single sensor to serve the user in the actual environment. It often requires multiple sensors to collect data in parallel and ensure the integrity and reliability of the user's data. As a result, data aggregation technology arises at the historic moment and becomes a key data processing technology in WSNs.

In order to improve the efficiency of sensor nodes and extend the life cycle of WSNs, data aggregation technology arises at the historic moment. The core idea of data aggregation technology is to process the original data of the different user nodes in WSN and reduce the data transmission without affecting the result. Thus, the purpose of prolonging the network life cycle and reducing the communication overhead is achieved. In recent years, data aggregation privacy protection in WSNs has received extensive attention. The data aggregation privacy protection protocol is focused on. At present, the existing privacy preserving data aggregation schemes can solve data leakage problems in data aggregation process in different degrees. Encryption of data is the most common way to protect data privacy. WSN determines its key distribution management because of its particularity, and the choice of encryption scheme is different from that of the conventional traditional network. In addition to encrypting node data, non-encrypted methods can also be selected for achieving privacy protection, such as data disturbance technology, anonymous technology, and hidden technology. According to the different practical application environment, it is necessary to take the amount of communication, computing overhead and privacy protection into consideration when choosing the privacy protection scheme, so as to choose the most suitable scheme.

2 Literature review

Privacy protection in WSNs has been studied by many researchers at present. According to the previous research results, the privacy protection schemes in WSNs are classified: data-oriented privacy protection and context-oriented privacy protection.

For data-oriented privacy protection, the relevant protocols of privacy protection are mainly to protect the sensitive data by encrypting and hiding data, so as to avoid the privacy disclosure caused by data access.

Chen et al. put forward a bucket privacy protection technology, which combines the bucket protection mechanism of data privacy protection with data encryption technology [1]. Qian et al. applied a bucket privacy protection technology to multidimensional data query privacy protection, and used a one-dimensional data-oriented privacy protection technology to the attribute value protection of multidimensional data [2]. He et al. proposed the concept of location privacy of source nodes. Source nodes are all important nodes in the application environment of many WSNs, so the location of its location is also very important [3]. Zhang et al. proposed a single path routing scheme, which combined random walk and single path routing, and used a single path routing algorithm to replace the second phases in the phantom flooding scheme. On the privacy

protection of base station location information, Zhang put forward a fixed communication amount of base station location privacy protection scheme. In this scheme, after the source node sends information, the packet is transmitted to the base station according to a fixed path, so the communication amount of each node is fixed, and the enemy cannot analyze the location of the base station according to the communication amount [4]. Li proposed a multi-base station protection scheme, which used pseudo base stations to establish pseudo communication to play the role of puzzling the enemy, thus protecting the location information of the real base station [5].

Data aggregation privacy protection in WSNs belongs to data-oriented privacy protection. In the era of big data, data-oriented privacy protection is one of the hottest research directions.

Joly et al. proposed the privacy protection of data aggregation based on the SMART protocol, which is based on the secure multi-party computing technology [6]. Using slicing technology, the sensor node data is divided into pieces, and the remaining slices are encrypted to the aggregator after the remaining pieces are retained. After a period of time, the nodes decrypt and aggregate all the data slices received. In addition, nodes can segment and exchange data, and the privacy protection is better. The disadvantage of the scheme is that a large amount of information interaction is required in the process of aggregation, resulting in large network traffic and poor data integrity. Subsequently, Dai improved the CPDA protocol and optimized the computation cost and communication overhead. In the proposed security scheme based on key disturbance, the base station shares a key with each sensor node in the proposed scheme. The data uploaded is composed of the original data and the key of the node, and the privacy protection of the original data and the aggregated data is realized [7]. In the data aggregation privacy protection protocol proposed by Arora et al., the base station and other nodes are unable to obtain the privacy data of a single node [8]. This protocol can tolerate partial data loss and communication failure. Zahurul et al. proposed a data privacy protection scheme in WSNs based on P-function sets. On the one hand, it can overcome the impact of data loss, and on the other hand, it can reduce the communication overhead [9]. The application environment of somatic area network in WSNs is studied. Firstly, the collected human health sensitive data is lossless compressed, and then the compressed sensitive data is embedded into ordinary data through data hiding technology to realize the privacy protection of sensitive data.

Olofsson et al. proposed a EPPA scheme. The aggregator in this scheme is semi-honest. In order to keep the user's privacy in the process of aggregation, the EPPA scheme uses the homomorphic encryption Paillier encryption technology to realize privacy protection in data aggregation. At the same time, the scheme is to transform the multidimensional data into a cipher-text form to realize the privacy protection of multidimensional data. They also put forward PDAFT scheme, which used homomorphic encryption Paillier encryption technology to encrypt user privacy data. Different from the EPPA scheme, in this scheme, the aggregated cipher cannot get the privacy data of a single user even in the control. In addition, the scheme also supports error tolerance and allows user data lost and error of misplaced server [10].

3 Method

A new protocol polynomial data aggregation (PRDA)+ is proposed. The data privacy protection is realized through the data disturbance technology. The integrity protection of the aggregated data is completed by the association between the two metadata. According to the least square polynomial curve fitting principle, the data of the sensor nodes is fitted into order polynomial function, so as to reduce the communication cost by uploading polynomial coefficients instead of the original node data.

3.1 System model

In WSNs, the network is divided into multiple clusters, each cluster is formed by a cluster head node (aggregator) and multiple sensor nodes. The cluster nodes can communicate with each other and upload the obtained sensing data to the aggregator. The aggregator can communicate with the adjacent aggregators, and finally, through multiple hops method, the aggregated data is sent to the base station.

Collection and aggregation of sensor nodes data: in the PRDA+ protocol proposed here, the sensor nodes collect and upload data. Similarly, the aggregator also collects and converts data to the base station at every fixed time. At the beginning of the data aggregation session, each aggregator broadcasts a message to notify members of its cluster to collect data. Each sensor node receives the message and begins to read the node data. The data aggregator requires the sensor node to collect and upload the data within a certain period of time. First, the nodes use n data to arrive at polynomial function by least square method. Then, the node uploads the coefficients of the polynomial function to the aggregator end instead of uploading the original node data. Next, the aggregator receives all the polynomial coefficients and it is polymerized within a certain period of time. Similarly, the coefficients are sent to the base station in the form of polynomials.

Polynomial representation of sensor node data: each sensor node preloads the curve fitting algorithm, and the data collected by sensor nodes are correlated. n is the sensor data set of sensors, and m is the order of polynomial functions. The polynomials are aggregated and expressed as:

$$D_{agg}(x) = \sum_{s} f_s(x) = \sum_{j} \left[\left(\sum_{s} a_{sj} \right) x^j \right]. \tag{1}$$

Key management: in the PRDA protocol, a random number generator is provided to the base station and each sensor node to obtain correct aggregated data, and through the random number generator (PRNG) of the synchronous base station, the random number as the same as node can be obtained. However, not all sensor nodes will participate in all processes of the aggregation session. Therefore, the PRGN of sensor nodes cannot generate random numbers, and the base station needs to synchronize PFNG multiple times to get the matched random number, and the result will lead to low aggregation efficiency.

The PRDA+ protocol proposed solves the above problem well. Before the network deployment, a random number generator is installed at the base station and all sensor nodes, but each sensor node shares the private key with the base station and becomes nodes $S_{bs,j}$ and $S_{bs,j}^{n}$. In the initialization phase of the aggregated data, a random number generator with its own private key is installed at each node, and each private key seed will be deployed in the node and shared with the base station at the beginning, so the base station knows the secret key of all nodes.

Attack model: in WSNs, the opponent can physically compromise the sensor nodes and get their private information. Therefore, an adversary can perform many kinds of attacks, including denial of service attacks, eavesdropping attacks, false data injection, modification, forgery or replay attacks. The PRDA protocol is designed to mitigate attacks on data privacy, so a simple sequence number method is used to prevent forgery attacks. Detection modification, false data entry and forgery require a wide range of monitoring mechanisms. Denial of service attacks blocking wireless channels cannot be effectively prevented in WSNs. The sensor node adds the message authentication code to the packet. Therefore, for data integrity attacks, denial of service attacks and replay attacks are not involved in this scheme. Assuming that the opponent's calculation is not strong than notebook computer, it means that it is not feasible to calculate an opponent to attack a message authentication code or a pseudo-random number in a violent attack.

3.2 PRDA+ for integrity verification

The PRDA+ protocol proposed makes use of the clustering idea to polymerize, disturb the data of the original node through the method of data disturbance, and add random numbers to realize the privacy protection. The communication overhead is reduced by expressing the polynomial function of node data and uploading the coefficients of polynomials. Finally, the integrity verification is carried out according to the association between the two metadata.

The aggregator broadcasts a message to the members within its cluster to notify the aggregation session to begin. This message contains a data aggregation session ordinal d, enabling sensor nodes to synchronize their PRNG to generate random numbers that match the session ordinal d. When the sensor node in the cluster receives this message, it reads the n readings of the sensor in a time period, and stores the n reading in the cache with the node. Taking into account that the cache cannot be too large and consumes memory, the previous readings will be covered when the next reading is passed in. When the aggregation message of the aggregator is received, the node fits the n readings to the m-order polynomial through the least squares, and the random number generator for each node generates a random number of the d aggregation sessions based on the node's secret key seed, which is recorded as r_i^d. Two random numbers are used to encrypt polynomials and the polynomials encrypted is recorded as:

$$Conf_i^d(x) = \sum_{j=0}^{m} \left(a_{ij} + r_i^d\right) x^j . \tag{2}$$

$$Conf_i^{nd}(x) = \sum_{j=0}^{m} \left(a_{ij} + r_i^{nd}\right) x^j . \tag{3}$$

In the proposed scheme, the polynomial encrypted by random numbers is cryptic, because the random number in this scheme is generated by the secret key seed shared by the base station and sensor nodes through PRNG. In addition to the base station and the sensor nodes, the outside world cannot obtain the secret key seed. Therefore, the encrypted polynomial, after aggregation, can completely achieve privacy protection.

Each sensor node sends the encrypted coefficients and the identifier (ID) of the node to the aggregator terminal. The aggregator performs a polymerization in a certain period of time to ensure that all sensor nodes transmit the data in the form of encrypted polynomials.

The base station receives aggregated data and lists from the aggregator end. First, by querying the ID list, it is known which nodes are involved in the aggregation. Then, the base station generates its corresponding random number according to its secret key seed, and the key seed and random number is shared by the node and the base station. The base station can get the decrypted aggregation data from the random number according to the encrypted aggregate value passed by the aggregator end, as shown in Formulas (4) and (5):

$$M_{agg}^{d}(x) = \sum_j \left[\left(\sum_s \left(a_{sj} + r^d - r^d\right) \right) x^j \right] \sum_j \left[\left(\sum_s a_{sj} \right) x^j \right]. \tag{4}$$

$$M_{agg}^{nd}(x) = \sum_j \left[\left(\sum_s \left(a_{sj} + r^{nd} - r^{nd}\right) \right) x^j \right] \sum_j \left[\left(\sum_s a_{sj} \right) x^j \right]. \tag{5}$$

After the base station obtains the decrypted aggregated data, the two are compared to verify the integrity. If the two values are equal, the data integrity of the aggregation can be verified. Conversely, the integrity of the data cannot be verified, and the aggregation is invalid.

4 Performance evaluation

The PRDA+ protocol proposed not only realizes the privacy protection of data aggregation, but also replaces the original node data by uploading the polynomial coefficients, which greatly reduces the communication overhead. By adding the secret key seed in the base station and the sensor node, the corresponding random number is generated and encrypted, and two encrypted aggregated data are uploaded to the aggregator. The base station decrypt two aggregated data and compare the two values. If the two values are equal, the data integrity verification is passed. Therefore, this scheme not only reduces the communication cost, but also obtains more accurate data and achieves integrity verification. In the Matlab simulation environment, the algorithm can be operated and a WSN with 200 sensor nodes randomly distributed in the 400m*400m

area is set up. The effective range of data transmission is 50m, and the data transmission rate is 1Mbps. It is assumed that the network is divided into N clusters and there are n sensor nodes in each cluster, and the data set is uploaded to the cluster head nodes because of the perceptual nodes. Therefore, the average selection for the cluster head to the base station is L for a hop communication. Assuming that each regression polynomial has m data items, the length of each coefficient is e, and the ID data length of the node is 4, then the communication overhead of the perceived node in the process of data fusion is as follows:

$$T_{i,t} = N * \left(1_{id} + e * m\right) * n . \tag{6}$$

For cluster head nodes, when each coefficient length is determined, the coefficient length of the aggregated polynomial is determined by the number of the aggregated polynomials, so the communication overhead of the cluster head nodes is:

$$T_{DA,t} = N * \left(1_{id} + e * m\right) * n * L . \tag{7}$$

Therefore, in the data aggregation process, the total communication overhead is:

$$T_{all} = N * n\left(1_{id} + e * m\right) * \left(L + 1\right). \tag{8}$$

4.1 Privacy protection analysis

Through two aspects, the privacy protection of PRDA+ protocol is analyzed.

Hypothesis 1: the enemy A eavesdrops the data transmitted by the sensor node, but it does not compromise the node.

It is proved that because the enemy does not compromise the node, it is impossible to know the secret key deployed at the sensor node at the initial stage. Without the secret key, the enemy cannot obtain the random number of the seeds generated according to the key, and cannot know the value of the current aggregation number d. Therefore, even if the enemy gets the data of the node through the eavesdropping attack, the encrypted data cannot be decrypted, and the data privacy of the sensor node can be protected.

Hypothesis 2: the enemy A compromises the aggregator to get encrypted aggregate data.

It is proved that the enemy compromises the aggregator to obtain two encrypted polynomials, and they are $ConD_{agg}^{d}\left(x\right)$ and $ConD_{agg}^{id}\left(x\right)$, respectively. If we want to decrypt the polynomial, it is necessary to know the secret key of the sensor node and the random number generated by the secret with the key as the seed, and also know the secret key of the base station and the seeds generated by the secret key of the base station. In this scheme, the secret key is deployed at the initial stage of the scheme between the node and the base station, and the secret key can only be shared between the sensor and the base station, and cannot be known by others. Therefore, private keys and random numbers cannot be obtained. Even if the enemy A compromises the

aggregated data obtained by the aggregator, the encrypted aggregated data cannot be deciphered. Therefore, it can be concluded that the data privacy of the aggregator terminal can also be protected.

4.2 Communication overhead

Matlab is used to do experiments, the PRDA+ protocol proposed here is compared with the PRDA protocol, and the impact of the PRDA+ protocol and the PRDA protocol on the communication overhead in the case of different m/n values is calculated, as shown in Figure 1.

The experimental results show that the data communication of the two protocols increases with the ratio increasing, and the data communication of the PRDA+ protocol is a little larger than that of the PRDA protocol under the same m/n ratio. However, the PRDA+ protocol ensures the integrity of the data through the correlation characteristics between the data, and the PRDAA protocol has not been implemented. As mentioned above, a slightly larger communication cost is inevitable, but it is not very different from the original one. Therefore, the scheme proposed here has certain superiority.

Similarly, the experiment also verifies the impact of cluster size on communication overhead, as shown in Figure 2.

The experimental results show that, with the gradual increase of clusters, the amount of data communication gradually decreases. The possible reasons are as follows: with the increase of the cluster, the number of sensor nodes in the cluster range will also increase. If the session is aggregated and the sensor nodes are uploaded simultaneously, it will inevitably result in the crowd. During this period, some data will inevitably be lost because of collision and the aggregation session will be missed, resulting in the communication amount of data reduced.

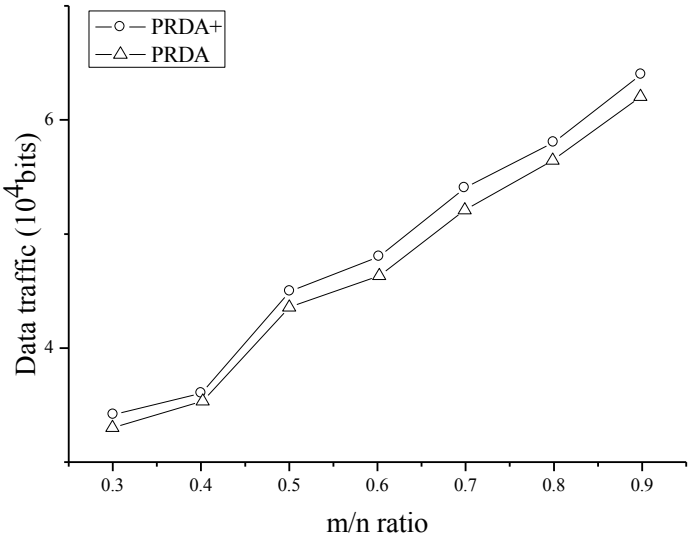

Fig. 1. The relationship between the ratio of m/n and the amount of data communication

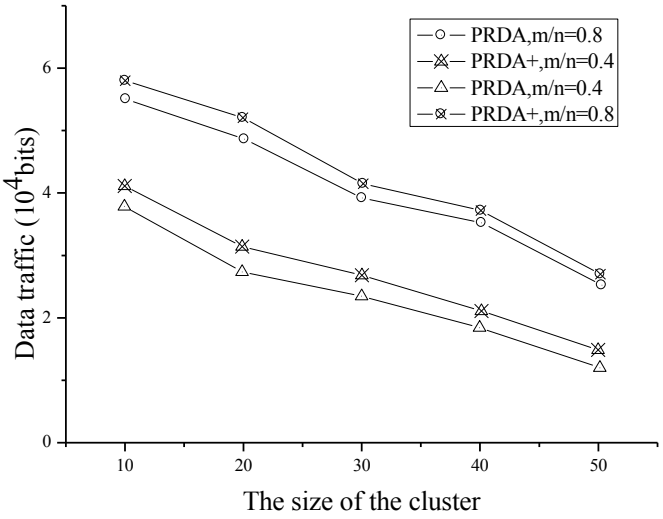

Fig. 2. The relationship between the size of a cluster and the amount of data communication

4.3 Data accuracy

The relationship between the accuracy of data and the ratio of m/n can be seen through a set of experiments. The experimental results, as shown in Figure 3 and Figure 4, show that the ratio of m/n is proportional to the accuracy of the data. With the increase of the ratio of m/n, the accuracy of the data is gradually increased. The reason is that the larger the ratio of m/n, the greater the m value, that is, the greater the order of polynomials, the more accurate the polynomial fitting the raw node data.

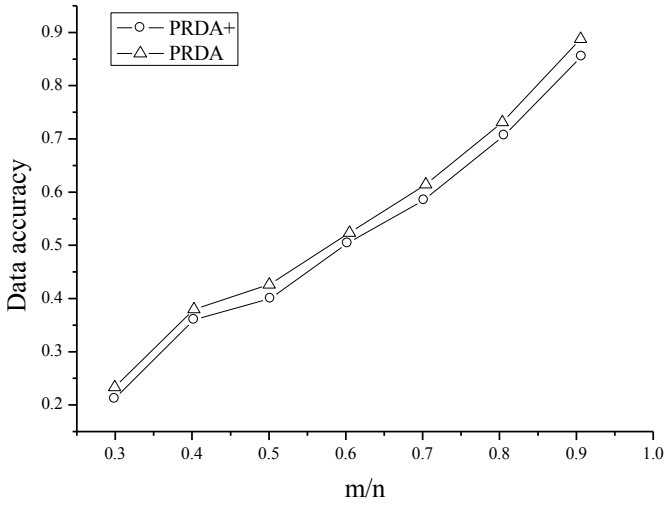

Fig. 3. The relationship between the ratio of m/n and the accuracy of data

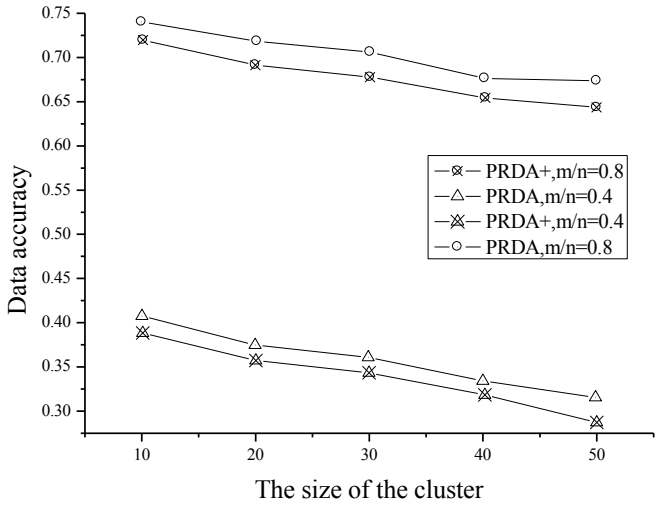

Fig. 4. The relationship between the size of a cluster and the accuracy of the data

The experiments show that the size of the cluster is inversely proportional to the accuracy of the data. The larger the cluster, the smaller the accuracy of the data. The reason is that the larger the cluster, the easier the aggregator to generate congestion to cause partial data loss. Therefore, the transmission data in the sensor nodes cannot be aggregated all, resulting in a decrease in accuracy.

4.4 Data integrity

Our PRDA+ protocol not only guarantees privacy protection in the process of data aggregation, but also provides integrity validation. A set of experiments will be used to show that our scheme provides better data integrity protection, as shown in Figure 5. The iCPDA protocol proposed by He is selected for the comparison with the scheme proposed here. From the experimental diagram, it can be seen that the proposed scheme can still guarantee the integrity of the data very well when the attack probability of the cluster head increases.

The experimental results show that with the increase of the probability of attack of cluster heads, the scheme proposed here has more advantages. The random number generated by the secret key seeds shared by the base station and the sensor nodes is decrypted. The two polynomials obtained are related, so it can effectively verify the integrity of the data.

The PRDA+ protocol proposed first synthesizes the n sensor data by the least square method to synthesize polynomial functions. Then, the node uploads the coefficients of the polynomial function to the aggregator end instead of uploading the original node data. Therefore, the amount of traffic is greatly reduced, and the scheme has more advantages for lightweight data. The scheme is different from other schemes in the key management. In the initialization stage, the random number generator is deployed in the sensor node and the base station, and the private key is allocated. The random num-

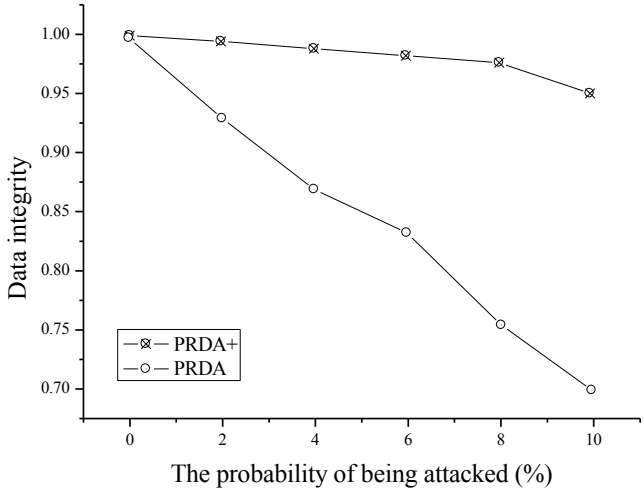

Fig. 5. Data integrity comparison

ber generator generates the corresponding number of machines according to the private key of each node. The polynomial coefficients synthesized by the sensor nodes are encrypted by the random number generated by their own private key seed, and then the base station and its shared secret key seed are generated by a random number to be encrypted. As a result, two encrypted polynomials can be uploaded to the aggregator end, and the aggregator end, after collecting the data of all the nodes, send the aggregated data and the corresponding node ID to the base station. The base station decrypts two corresponding random numbers according to the private key seed of its own and the node, and obtains two aggregated data for comparison. If two aggregated data are consistent, data integrity verification is adopted. To sum up, the proposed PRDA+ protocol can effectively protect the privacy of aggregated data, reduce communication overhead and achieve data integrity protection.

5 Conclusion

With the rapid development of WSN, its security problem has also received wide attention. Data aggregation privacy protection in WSNs is studied. The PRDA+ protocol is proposed, which realizes the privacy protection of data, reduces the communication overhead, and realizes integrity verification. It belongs to the data aggregation privacy protection technology based on polynomial regression.

The PRDA+ protocol proposed uploads the coefficients of polynomial functions to the aggregator terminal instead of uploading the original node data, and the communication amount is greatly reduced. In the key management, the scheme is in the initialization stage, the random number generator is deployed in both the sensor node and the base station, and the private key is allocated. The random number generator encrypts the data according to the random number of each node to generate the corresponding random number of the seed. The base station decrypts two corresponding random

numbers according to the private key seed of itself and the node, and compares the two aggregated data to realize the integrity verification. In summary, the PRDA+ protocol proposed can effectively protect the privacy of aggregated data, reduce communication overhead and achieve data integrity protection.

6 References

[1] Chen, J., Ma, H., & Zhao, D. (2017). Private data aggregation with integrity assurance and fault tolerance for mobile crowd-sensing. Wireless Networks, 23(1): 131-144. https://doi.org/10.1007/s11276-015-1120-z

[2] Qian, J., Qiu, F., Wu, F., Na, R., Chen, G., & Tang, S. (2017). Privacy-preserving selec-tive aggregation of online user behavior data. IEEE Transactions on Computers, 66(2): 326-338.

[3] He, D., Kumar, N., & Lee, J. H. (2016). Privacy-preserving data aggregation scheme against internal attackers in smart grids. Wireless Networks, 22(2): 491-502. https://doi.org/10.1007/s11276-015-0983-3

[4] Zhang, Y., Chen, X., Li, J., Wong, D. S., Li, H., & You, I. (2016). Ensuring attribute privacy protection and fast decryption for outsourced data security in mobile cloud computing ☆. Information Sciences, 379.

[5] Li, Y., Dai, W., Ming, Z., & Qiu, M. (2016). Privacy protection for preventing data over-collection in smart city. IEEE Transactions on Computers, 65(5): 1339-1350. https://doi.org/10.1109/TC.2015.2470247

[6] Joly, Y., Dyke, S. O. M., Knoppers, B. M., & Pastinen, T. (2016). Are data sharing and privacy protection mutually exclusive? Cell, 167(5): 1150-1154. https://doi.org/10.1016/j.cell.2016.11.004

[7] Dai, W., Qiu, M., Qiu, L., Chen, L., & Wu, A. (2017). Who moved my data? privacy pro-tection in smartphones. IEEE Communications Magazine, 55(1): 20-25. https://doi.org/10.1109/MCOM.2017.1600349CM

[8] Arora, V. K., Sharma, V., & Sachdeva, M. (2016). A survey on leach and other's rout-ing protocols in wireless sensor network. Optik, 127(16): 6590-6600. https://doi.org/10.1016/j.ijleo.2016.04.041

[9] Zahurul, S., Mariun, N., Grozescu, I. V., Tsuyoshi, H., Mitani, Y., & Othman, M. L., et al. (2016). Future strategic plan analysis for integrating distributed renewable genera-tion to smart grid through wireless sensor network: malaysia prospect. Renewable & Sustainable Energy Reviews, 53: 978-992. https://doi.org/10.1016/j.rser.2015.09.020

[10] Olofsson, T., Ahlén, A., & Gidlund, M. (2016). Modeling of the fading statistics of wireless sensor network channels in industrial environments. IEEE Transactions on Signal Processing, 64(12): 3021-3034. https://doi.org/10.1109/TSP.2016.2539142

7 Authors

Bohua Guo works as associate professor at Shandong Binhai University, Shandong, China.

Yanwu Zhang works as associate professor at Shandong Binhai University, Shandong, China.

Article submitted 07 September 2018. Resubmitted 18 October 2018. Final acceptance 25 October 2018. Final version published as submitted by the authors.