Digital Communications and Networks 2 (2016) 122–129

HOSTED BY

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.elsevier.com/locate/dcan

Scalable privacy-preserving big data aggregation mechanism

Dapeng Wu *, Boran Yang, Ruyan Wang

School of Information and Communication Engineering, Chongqing University of Posts and Telecommunications, No. 2, Chongwen Road, Nan'An District, Chongqing City 400065, PR China

ARTICLE INFO

Article history:
Received 20 February 2016
Received in revised form
1 July 2016
Accepted 5 July 2016
Available online 15 July 2016

Keywords:
Big sensor data
Privacy-preserving method
Data aggregation
Gradient-based clustering

ABSTRACT

As the massive sensor data generated by large-scale Wireless Sensor Networks (WSNs) recently become an indispensable part of 'Big Data', the collection, storage, transmission and analysis of the big sensor data attract considerable attention from researchers. Targeting the privacy requirements of large-scale WSNs and focusing on the energy-efficient collection of big sensor data, a Scalable Privacy-preserving Big Data Aggregation (Sca-PBDA) method is proposed in this paper. Firstly, according to the pre-established gradient topology structure, sensor nodes in the network are divided into clusters. Secondly, sensor data is modified by each node according to the privacy-preserving configuration message received from the sink. Subsequently, intra- and inter-cluster data aggregation is employed during the big sensor data reporting phase to reduce energy consumption. Lastly, aggregated results are recovered by the sink to complete the privacy-preserving big data aggregation. Simulation results validate the efficacy and scalability of Sca-PBDA and show that the big sensor data generated by large-scale WSNs is efficiently aggregated to reduce network resource consumption and the sensor data privacy is effectively protected to meet the ever-growing application requirements.

1. Introduction

As countless sensor technology breakthroughs enable the widespread implementations of large-scale Wireless Sensor Networks (WSNs), the sensor data collected by individual sensors contributes to the huge overall data volume of the network [1]. The big sensor data becomes an indispensable part of 'Big Data', the collection, storage, transmission and analysis of the big sensor data attract considerable attention from researchers [2]. The WSN [3] consists of low-cost and high-performance sensor nodes responsible for objective monitoring and data reporting, and the sink for sensor data gathering, which is widely applied to health care [4], food processing, environment monitoring, and wildlife tracking, As indicated in [5], the daily activities of people and machines nowadays are monitored and measured by sensors in unimaginable ways, therefore the collection, storage, transmission and analysis of the big sensor data are inevitably faced with severe challenges.

Firstly, in numerous practical sensor network implementations such as forest fireproofing, senior healthcare and military surveillance, sensor data is privacy-sensitive and their security and privacy should be properly protected. In terms of the data privacy, the wireless medium, supporting the transmission of big sensor data, is generally exposed to severe privacy threats such as the data leakage and eavesdropping [6]. Especially for large-scale wireless sensor networks, the secure transmission of sensor data and the privacy of sensor nodes have to be effectively guaranteed [7]. Besides, the multi-hop transmission manner of sensor data in WSN enables the extra-long range data transmission while causing the "hot spot" problem for relay nodes close to the sink. Therefore, the resource restraints of sensor nodes should also be taken into account due to the heavy energy consumption of relay nodes, and the design of an efficient privacy-preserving method for big sensor data is crucial for the universal application of WSN.

Secondly, in networks with densely deployed sensors, sensor data is always spatially correlated and the unnecessary transmission of redundant data will aggravate the network resource consumption [8]. The data aggregation technology [9] is commonly employed to reduce the transmission size and times of sensor data in WSNs, by exploiting the correlation of sensor data collected by neighbor sensors. The data aggregation method is especially beneficial for large-scale WSN, because it can sufficiently aggregate the big sensor data, effectively avoid the transmission of unnecessary redundant data and accurately extract the data features for the sink node to make the right decisions.

Due to the importance of the aggregated big sensor data, the privacy of the aggregated results is extremely vulnerable without

* Corresponding author.
E-mail address: wudapengphd@gmail.com (D. Wu).
Peer review under responsibility of Chongqing University of Posts and Telecommunications.

http://dx.doi.org/10.1016/j.dcan.2016.07.001

employing any privacy-preserving technology, and interception of the unprotected aggregation results will severely harm the network owner. Therefore, privacy-preserving technology is essential for the big sensor data aggregation, which can substantially prevent the original sensor data or the aggregated results from attacking [10]. Furthermore, the scalability and energy efficiency [11] of privacy-preserving big data aggregation should also be considered.

Targeting the privacy-preserving requirements and the energy-efficient transmission of big sensor data, this paper designs a Scalable Privacy-preserving Big Data Aggregation (Sca-PBDA) method for WSN. Firstly, according to the pre-established gradient topology structure, sensor nodes in the network are divided into clusters with a Cluster Head (CH),an equal number of Cluster Members (CMs)and different numbers of auxiliary Cluster Heads (aCHs). By equally clustering the network, the privacy-preserving configuration and intra-cluster data aggregation methods are uniformly designed to achieve the further inter-cluster data aggregation. Besides, by adaptively allocating aCH for each cluster, the heavy energy consumption of CHs close to the sink can be balanced and the "hot-spot" problem can be solved. Secondly, sensor data is modified by each node according to the privacy-preserving configuration message received from the sink. After the configuration, the privacy-preserving data is sent to the CH for the intra-cluster data aggregation. During the inter-cluster data transmission process, the aggregated privacy-preserving data is re-aggregated at the relay CHs, until they reach the sink. Eventually, the sink recovers the aggregation result from the received privacy-preserving aggregation data. With the proposed Sca-PBDA, the big sensor data generated by large-scale WSN is efficiently aggregated to reduce the network resource consumption and the sensor data privacy is effectively protected to meet the ever-growing application requirements. The major contributions of this paper are summarized as follows:

- Agradient-based equal network clustering method is proposed in this paper to reasonably determine the network topology, according to the estimated node energy consumption. With the proposed clustering method, the identical number of CH and CMs supports the uniform privacy-preserving configuration and further inter-cluster data aggregation process, which meets the scalability requirements of the big sensor data collected by large-scale WSNs.
- A scalable privacy-preserving data aggregation method is further designed in this paper to provide the simple privacy-preserving data configuration and scalable intra- and inter-cluster data aggregation. Especially, the effectively protected big sensor data is parallel aggregated at each CH, and relay CHs also perform aggregation operations on the received privacy-preserving aggregation data to further reduce resource consumption, by which the processing load of the sink is shared among CHs.

The rest of this paper is organized as follows. The network gradient establishing procedure is firstly completed and the energy consumption of two different types of nodes is analyzed to further design a reasonable clustering method in Section 2. Section 3 introduces the privacy-preserving data configuration and the intra-cluster data aggregation methods. Section 4 further proposes the inter-cluster data aggregation to meet the scalability requirements of big sensor data and the aggregation result recovery method by the sink is introduced in Section 5. Subsequently, the settings of the simulations are given and the performance of Sca-PBDA is analyzed in Section 6. Section 7 recommends the related works. Lastly, the conclusion is given in Section 8.

2. Clustering method

2.1. Gradient establishment

In actual large-scale WSN scenarios, nodes are always randomly distributed and are classified into two categories: nodes generate sensor data according to the monitored objectives, and the sink that gathers and processes the reported big sensor data. Clustering is an efficient network topology control method especially for large-scale WSNs, which selects CHs among each cluster to perform further aggregating and relaying operations on big sensor data and establishes a hierarchical network structure. To reasonably cluster the network, the Gradient Establishing (GE) message is firstly broadcast by the sink in the circular sensing field, by which other randomly distributed sensor nodes determine their distances to the sink. For instance, when receiving the broadcasted GE, sensors within the communication range of the sink will update the hop count field of GE after obtaining their gradient value 1 (the gradient value of the sink is 0). Adding the hop count of GE by 1, these sensors rebroadcast the updated GE after a certain delay. Besides, sensor nodes determine their gradient value, namely the hop count distance from the sink, only according to the first received GE message. By repeating this procedure, the GE message is broadcasted, updated and rebroadcasted within the network to effectively determine the gradient value of every sensor, which further supports the energy consumption analysis and reasonable clustering method.

Generally, resource and capability restrained sensor nodes are not wired and equipped with rechargeable batteries, and CHs may be selected among them to aggregated big sensor data for the reduced transmission size and times and energy consumption. However, packets generated by nodes far away from the sink also have to be forwarded by the relay CHs to reduce the communication cost. Thus the energy consumption of CHs closer to the sink is much heavier, and should be considered when designing algorithms especially for big sensor data. Mechanisms exploiting CH rotating or unequal clustering technologies are optional solutions for the so-called 'hot-spot' problem. However, for the scalability of the uniform privacy-preserving intra-cluster aggregation method, the identical cluster composition of CH and CMs with a different number of a CHs is employed in this paper. To reasonably allocate aCHs for each cluster, the energy consumption of CHs with different gradient values needs to be analyzed.

2.2. Energy analysis

Obviously, the energy consumption of CHs is negatively correlated with their gradient values due to the relaying function and directly proportional to the number of CMs due to the number of received data packets. Mathematically, the total energy consumption of CH_i with gradient value i consists of 3 components as shown in Eq. (1), where $EC_{Receiving}$ is the energy consumption of receiving datapackets sent from its CMs and other CHs, $EC_{Aggregating}$ is the energy consumption of aggregating these received datapackets, and $EC_{Transmitting}$ is the energy consumption of forwarding the aggregated datapackets.

$$EC_{CH_i} = EC_{Receiving} + EC_{Aggregating} + EC_{Transmitting} \quad (1)$$

Specifically, $EC_{Receiving}$, $EC_{Aggregating}$ and $EC_{Transmitting}$ are denoted by Eq. (2), where CS is the cluster size, namely the number of CH and CMs within each cluster, L_{packet} is the data packet length (Byte), N_{RP}^{i+1} is the average number of data packets received from CHs with gradient value $i+1$ by CH_i. In addition, EC_{Rx}, EC_{DA}, EC_{Tx} and EC_{amp} denote the energy consumption of receiving, aggregating, transmitting and amplifying each data byte, d is the transmission

distance, and $\alpha \geq 2$ is the transmission power loss coefficient, which increases as the appearance of obstacles in the transmission path and equals 2 in free space.

$$\begin{cases} EC_{Receiving} = (CS-1)\cdot L_{packet}\cdot EC_{Rx} + N_{RP}^{i+1}\cdot L_{packet}\cdot EC_{Rx} \\ EC_{Aggregating} = (CS-1)\cdot L_{packet}\cdot EC_{DA} + N_{RP}^{i+1}\cdot L_{packet}\cdot EC_{DA} \\ EC_{Transmitting} = L_{packet}\cdot\left(EC_{Tx} + d^{\alpha}\cdot EC_{amp}\right) \end{cases} \tag{2}$$

For each CH_i, they receive N_{RP}^{i+1} aggregated data packets on average from CH_{i+1}, which is calculated as shown in Eq. (3). In Eq. (3), ρ is the sensor node density and r_g is the average gradient radius, which reflects the coverage of each gradient.

$$N_{RP}^{i+1} = \frac{\left[\pi(i+1)^2 r_g^2 - \pi i^2 r_g^2\right]\rho}{\left[\pi i^2 r_g^2 - \pi(i-1)^2 r_g^2\right]\rho} = \frac{2i+1}{2i-1} \tag{3}$$

Therefore, the energy consumption of CH_i is reduced to Eq. (4).

$$EC_{CH_i} = \left(CS + N_{RP}^{i+1} - 1\right)\cdot L_{packet}\cdot\left(EC_{Rx} + EC_{DA}\right) + L_{packet}.\left(EC_{Tx} + d^{\alpha}\cdot EC_{amp}\right) \tag{4}$$

However, CMs on the network only have to send their privacy-preserving sensor data to the corresponding CHs and are not responsible for receiving sensor data and performing aggregation operations. Therefore, the total energy consumption of a CM is calculated as shown in Eq. (5).

$$EC_{CM} = L_{packet}\cdot\left(EC_{Tx} + d^{\alpha}\cdot EC_{amp}\right) \tag{5}$$

Obviously, the number of a required CHs for clusters is estimated according to the proportion of the energy consumed by CH_i to the energy consumed by a CM, as shown in Eq. (6).

$$N_{BCH_i} = \frac{EC_{CH_i}}{EC_{CM}} - 1 = \frac{\left(CS + N_{RP}^{i+1} - 1\right)\cdot\left(EC_{Rx} + EC_{DA}\right)}{\left(EC_{Tx} + d^{\alpha}\cdot EC_{amp}\right)} = \frac{\left(CS + \frac{2}{(2i-1)}\right)\cdot\left(EC_{Rx} + EC_{DA}\right)}{\left(EC_{Tx} + d^{\alpha}\cdot EC_{amp}\right)} \tag{6}$$

2.3. Clustering process

By adaptively allocating aCH for clusters, the energy consumption of the relay CHs can be effectively shared to avoid the energy depletion of relay CHs closer to the sink, which might cause network partitioning and interrupt communication. With the identical cluster composition of CH and CMs, the inter-cluster privacy-preserving data aggregation can be further achieved to meet the scalability requirements of big sensor data.

During the network initialization stage, the sink establishes the network gradient by broadcasting the GE message. After the gradient establishes procedure, sensor nodes with the gradient value i select themselves as CHs with probability $P_{CH}^{i} = \frac{1}{\left(CS + N_{BCH_i}\right)}$. Besides, the competition mechanism is employed to ensure the uniqueness of a CH within its communication range. Subsequently, other sensor nodes send the joining message request to join a cluster as CMs or aCHs. The network structure is illustrated by Fig. 1.

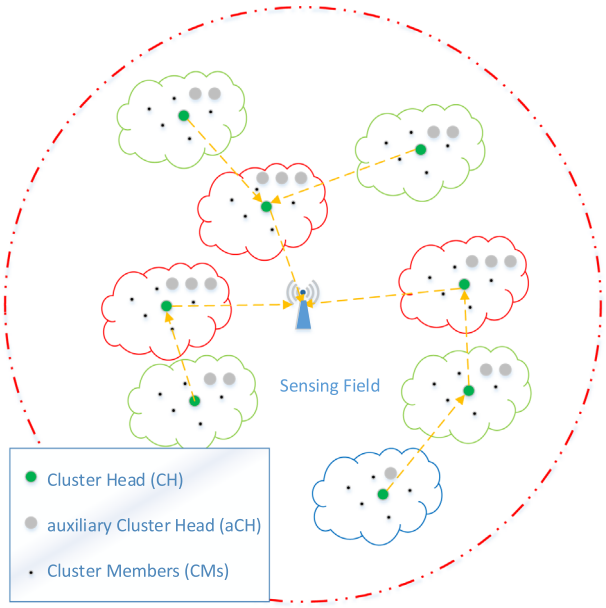

Fig. 1. Network structure.

3. Intra-cluster privacy-preserving data aggregation

3.1. Privacy-preserving positions

Based on the clustered network structure, the sink first determines the Global True value Position Set (*GTPS*), which is randomly generated to tag the true sensor data values. Besides, $GTPS \subset I$ and I is the data index set for the sensor data generated by each node, which indicates that the privacy-preserving sensor data consists of true sensor data values and camouflage filling values. With the camouflage filling values, the privacy of true sensor data values is effectively guaranteed and the aggregation operation can be performed. Obviously, the size of I is crucial and determinant of the privacy-preserving performance, because a larger $|I|$ can provide space for sensor nodes to fill more camouflage values. However, the increasing $|I|$ will inevitably cause additional communication overhead. According to the privacy-preserving performance requirements of actual applications, I is determined by the sink. In a large-scale WSN without clustering technology, the size of *GTPS* can be extremely huge, which consequently causes the increase of I, the size of transmitted privacy-preserving data packets and the communication energy consumption. To efficiently achieve the privacy-preserving data aggregation on big sensor data, the clustered network topology is employed by Sca-PBDA and CS is set as the size of *GTPS*, namely $|GTPS| = CS$, which substantially reduces the $|I|$ and the size of privacy-preserving data packets. Therefore, Sca-PBDA guarantees excellent energy efficiency and favorable scalability.

Furthermore, the sink allocates Node Private Position Set (*NPPS*) and Node True value Position Set (*NTPS*) for each sensor where $NTPS \subset GTPS \subset NPPS \subset I$ and $\cup_{i=1,\cdots,CS} NTPS_i = GTPS$. Especially, for MAX and MIN aggregation functions, $|NTPS| = 1$ generally holds. In the proposed Sca-PBDA, the MAX aggregation function is employed and *NTPS* tags the position of the true sensor data value for each node. Therefore, the *NTPS* assigned for each node is different and every node only knows the position of their own true sensor data value, while the sink knows the position set of all true sensor data values, namely $GTPS = \cup_{i=1,\cdots,CS} NTPS_i$. *NPPS* tags the positions for each node to place the true sensor data value and the restricted camouflage values. Besides, each node in a given cluster only knows its own *NPPS*, which prevents nodes from obtaining the *GTPS*, namely $GTPS \subset NPPS$. By reasonably configuring *NTPS* and *NPPS* for each node, the further privacy-preserving camouflage filling is effectively achieved.

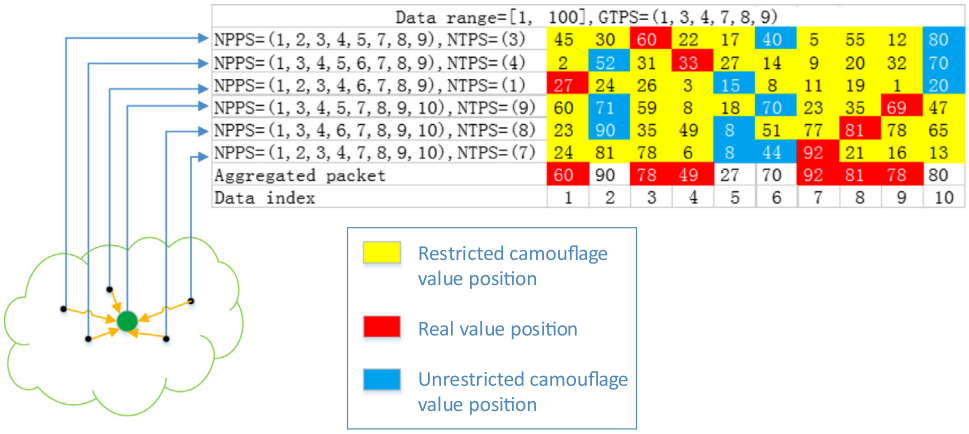

Fig. 2. Intra-cluster privacy-preserving data aggregation. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

3.2. Camouflage filling

Before the intra-cluster data aggregation, the sink disseminates the configured *NTPS* and *NPPS* for each sensor node. Due to the equal *CS*, every cluster employs the similar privacy-preserving camouflage filling method, which further facilitates the inter-cluster privacy-preserving data aggregation. As shown in Fig. 2, according to the received privacy-preserving data configuration message *NTPS*and *NPPS*, true sensor data value, restricted camouflage values and unrestricted camouflage values are placed in the red, yellow and blue positions respectively. The restricted camouflage value position set (*RCPS*) is derived from $RCPS = NPPS - NTPS$, and their values are randomly generated within the numerical range, which is restricted by the aggregation function, as shown in Eq. (7), where $d^{restricted}$ is the restricted camouflage values and d^{real} is the true sensor data value.

$$\begin{cases} d^{restricted} < d^{real} & for\ MAX\ aggregation \\ d^{restricted} > d^{real} & for\ MIN\ aggregation \end{cases} \tag{7}$$

With this numerical range, the aggregated value of all true sensor data values within the given cluster, namely the intra-cluster maximum value, will not be covered by the restricted camouflage values in the same positions during the aggregation process.

The Unrestricted Camouflage value Position Set (*UCPS*) is derived from $UCPS = \overline{NPPS}$, and the values of the unrestricted camouflage data are randomly generated within the numerical range of sensor data. For instance, with the MAX aggregation function adopted, all yellow restricted camouflage values are smaller than the red true sensor data value. If attackers intercept the data packet without unrestricted camouflage values, by directly extracting the maximum value, the truesensor data value can be easily cracked. Therefore, by introducing blue unrestricted camouflage data, the truesensor data value can be effectively protected and the cracking difficulty will be greatly improved. Furthermore, because the position sets of truesensor data values and unrestricted camouflage values are disjoint, the data aggregating process is not affected. The abovementioned *NPPS* consists of *NTPS* and *RCPS*, and the restricted camouflage data ensures the effective aggregation of truesensor data values and protects the *GTPS*, whereas the unrestricted camouflage data guarantees the privacy of truesensor data.

3.3. Data aggregation

After completing the privacy-preserving data configuration, the protected data is sent to the CH. Due to the gradient-based clustering method, the data transmission between sensor nodes and the sink is achieved in the multi-hop manner, whereas the single-hop data transmission is employed between CMs and CHs [12]. When a CH receives the privacy-preserving data packets from all its CMs, the CH directly performs the MAX data aggregation operation on them, as shown in Eq. (8), where $Data_{Aggregated}$ is the aggregated privacy-preserving data and d_{ij} denotes the data value of ith position by node j.

$$Data_{Aggregated} = \bigcup_{i=1,\cdots,I} \max_{j=1,\cdots,CS} \left(d_{ij}\right) \tag{8}$$

$Data_{Aggregated}$ contains the maximum value of all sensor data values of the given cluster. However, the original sensor data of each node in $Data_{Aggregated}$ is probably covered by the restricted camouflage values of other nodes, and attackers also cannot recover the original sensor data of each node, even if they may intercept the aggregated privacy-preserving data. Therefore, big sensor data are parallel processed at CHs to achieve energy-efficient data aggregation and the privacy of sensor data and sensor nodes is also protected. Furthermore, the aggregated privacy-preserving data packets have to be forwarded by relay CHs with lower gradient values and reported to the sink.

4. Inter-cluster privacy-preserving data aggregation

During the inter-cluster data forwarding phase, employing the same data aggregation method is possible due to the identical cluster composition and similar privacy-preserving camouflage filling method. Besides, the inter-cluster privacy-preserving data aggregation further reduces the data transmission times while reserving the aggregated result. During the intra-cluster data aggregation phase, the original sensor data of each node is probably covered by the restricted camouflage values of other nodes, therefore the main purpose of the intra-cluster data aggregation is to reserve the maximum sensor data value of this cluster. Obviously, the sink can recover this maximum sensor data value by scanning positions tagged by *GTPS*. With the goal of obtaining the global aggregated result, the aggregated privacy-preserving data is re-aggregated at the relay CHs, as shown in Eq. (9), where $Data_{reAggregated}$ is the re-aggregated privacy-preserving data and N_{RP} denotes the number of received privacy-preserving aggregation data packets.

$$Data_{reAggregated} = \bigcup_{i=1,\cdots,I} \max_{j=1,\cdots,N_{RP}} \left(d_{ij}\right) \tag{9}$$

After the inter-cluster privacy-preserving data aggregation, the $Data_{reAggregated}$ contains the maximum sensor data value of all

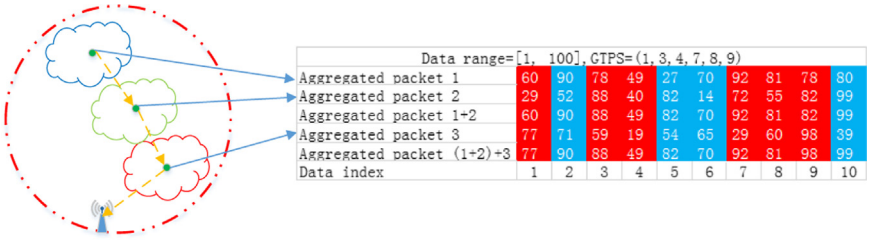

Fig. 3. Inter-cluster privacy-preserving data aggregation.

sensors on this link. As shown in Fig. 3, the given link includes 3 clusters and the sink, and the outermost CH_3 first sends its aggregated intra-cluster privacy-preserving data to CH_2 for the further re-aggregation. The re-aggregation packet generated by CH_2 contains the maximum sensor data value of cluster 3 and 2. Similarly, the re-aggregation packet received by the sink contains the maximum sensor data value of all clusters on this link. More importantly, the pre-established network gradient topology supports the inter-cluster transmission of aggregated privacy-preserving data packets from the outermost cluster to the sink. During the inter-cluster privacy-preserving data aggregation, CHs do not know the detailed contents of their own or the received intra-cluster privacy-preserving aggregation data, and only perform the simple re-aggregation operation on them, which effectively guarantees the privacy of transmitted big sensor data.

5. Recovery of the aggregated result

After receiving all aggregated privacy-preserving sensor data packets from CH_1s, the sink performs the last aggregation operation on them, scans the data index positions tagged by *GTPS* and takes the maximum value as the aggregated result, namely the global maximum value of the big sensor data. As shown in Fig. 4, because the *GTPS* is determined by the sink during the privacy-preserving data configuration phase, the recovery of the aggregated results is successfully achieved by scanning $Data_{global} = \bigcup_{i=1,\cdots,I} \max_{j=1,\cdots,N_{Cluster}} (d_{ij})$, where $Data_{global}$ is the global aggregated privacy-preserving sensor data packet and $N_{Cluster}$ is the total cluster number of CH_1s in the given WSN. Besides, *GTPS* is only kept by the sink and the other sensor nodes only obtain the allocated *NPPS* and *NTPS*. Therefore, even if the aggregated privacy-preserving big sensor data packets are intercepted by other nodes or attackers, they cannot successfully crack the aggregated results without knowing *GTPS*. The privacy of the big sensor data is efficiently guaranteed by the proposed Sca-PBDA, due to the low computational complexity compared with the cryptographic methods, and the privacy of the sensor nodes is also be ensured due to the intra- and inter-cluster data aggregation. Most importantly, with the designed gradient-based clustering method and inter-cluster privacy-preserving data aggregation, the energy-efficient parallel processing of the big sensor data is achieved and the scalability requirements of large-scale WSNs are met.

6. Simulation analysis

The Matlab simulation platform is adopted in this paper to validate the efficacy and scalability of Sca-PBDA. Firstly,we simulated the network topology established by Sca-PBDA. Secondly, the privacy performance of the proposed Sca-PBDA is analyzed and compared with those of traditional WSN privacy-preserving data aggregation mechanisms such as CPDA and KIPDA. Furthermore, the communication overheads of these privacy-preserving data aggregation mechanisms are analyzed and compared. Eventually, the network lifetimes of these mechanisms are simulated and analyzed.

As is seen in Fig. 5, the network topology is reasonably established by the proposed gradient-based equal clustering method, where the circled dots of various colors denote CHs of different gradients, the dots of various colors denote the CMs of different gradients, and the stars denote the CHs. Obviously, the clusters closer to the sink have more CHs according to our design, and the network is equally clustered.

To objectively verify the privacy-preserving performance of Sca-PBDA, the transmission processes of the position configuration messages are assumed to be privileged and cannot be intercepted and interpreted.

Let the crack probability of each link be q, when employing Sca-PBDA, the data leak probability of the given link is calculated by Eq. (10), which is defined as the probability that an attacker identifies the true sensor data from the unrestricted camouflage data when a given link is cracked.

$$P_{Sca\_PBDA}(q) = q \cdot \frac{1}{CS - GTPS + 1} \tag{10}$$

Obviously, attackers have to crack all links within a given cluster to decrypt the aggregated data. Given the cluster size *CS*, the data leak probability of the given cluster is calculated by Eq. (11), which indicates the probability that all links between the CH and CMs are cracked.

$$\mathbb{P}_{Sca\_PBDA}(q) = 1 - (1 - P_{Sca\_PBDA}(q))^{CS-1} \tag{11}$$

For CPDA, data leaks only occurs during the packet exchange

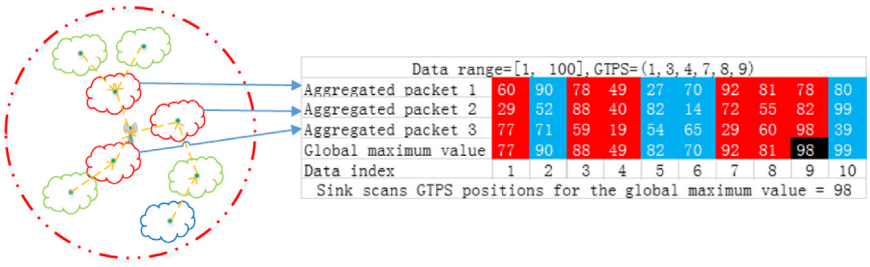

Fig. 4. Recovery of the aggregated result.

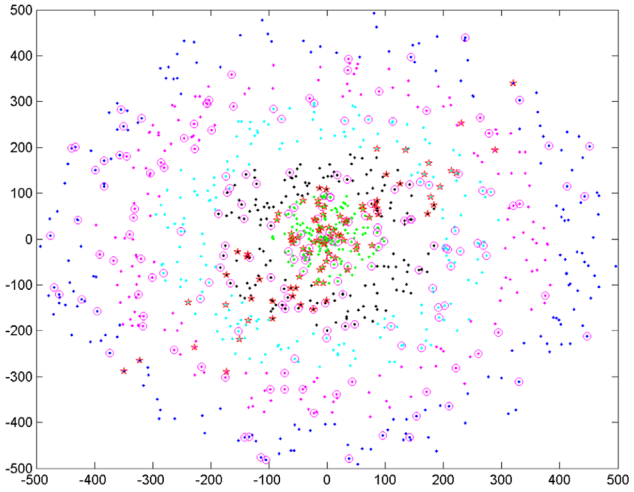

Fig. 5. Simulated network topology. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

process in a given cluster. Given cluster size CS, each CM sends $(CS-1)$ encrypted packets to the other $(CS-1)$ CMs for the privacy-preserving data aggregation. Attackers must intercept all the encryption keys to crack the privacy of a packet, therefore the data leak probability is calculated by Eq. (12), given crack probability of each link q. If an attacker cracks packets from every single node in a given cluster, the data security of the cluster is regarded as cracked.

$$\mathbb{P}_{CPDA}(q) = 1 - (1 - q^{CS-1})^{CS} \tag{12}$$

As is seen from Fig. 6, given a relatively small CS, the data encryption complexity of CPDA is low, therefore its data leak probability is higher when compared with Sca-PBDA. Under the scenario of larger CS, the data encryption complexity of CPDA is dramatically enhanced, which leads to better privacy performance when compared with the proposed Sca-PBDA. However, the larger CS of CPDA incurs extremely heavy communication and computation overheads, which greatly heightens the deployment cost and implemental difficulties of WSN applications. The proposed Sca-PBDA achieves favorable privacy performance under various cluster sizes.

Similar to other types of networks, data transmissions are the major contribution to the energy consumption in WSN [13]. Therefore, the communication overhead should be considered when quantifying the energy efficiency and determining the scalability of privacy-preserving data aggregation mechanisms.

During the data aggregation processes of Sca-PBDA and KIPDA, each CM receives 1 privacy-preserving configuration message via CH from the sink, and sends 1 privacy-preserving packet to the CH, therefore the communication overheads of each CM are both 2. Furthermore, during the data aggregation process of CPDA, each CM broadcasts 1 and receives $(CS-1)$ random seeds, then sends 1 and receives $(CS-1)$ encrypted packets, and eventually sends 1 encrypted packet to the CH for data aggregation, therefore the communication overhead of each CM is calculated, as shown in Eq. (13).

$$\begin{aligned} CommO_{CPDA} &= \big(1 + (N-1)\big)\cdot\frac{Length(Seed)}{Length(Packet)} \\ &\quad + \big((N-1) + (N-1) + 1\big) \\ &= 2N - 1 + \mathrm{N}.\frac{Length(Seed)}{Length(Packet)} \end{aligned} \tag{13}$$

The simulated communication overheads of 3 mechanisms are illustrated in Fig. 7. Although the number of transmitted packets is independent from the network size, the length of the transmitted packets increases along with the growing number of nodes in KIPDA. According to our design, the communication overhead of the proposed Sca-PBDA is substantially reduced due to the restrained data packet length and inter-cluster data aggregation process. In CPDA, the communication overhead increases along with the growing CS and is related to the length of the random seed. However, the length of the random seed is far smaller than that of the data packet, thus the extra communication overhead is insignificant. Due to the frequent intra-cluster communication to achieve the data privacy of CPDA, its communication overhead is much larger than the camouflage filling privacy-preserving methods, such as KIPDA and Sca-PBDA. More importantly, the communication overhead of Sca-PBDA smoothly increases with the growing CS, which signifies its scalability for large-scale WSNs.

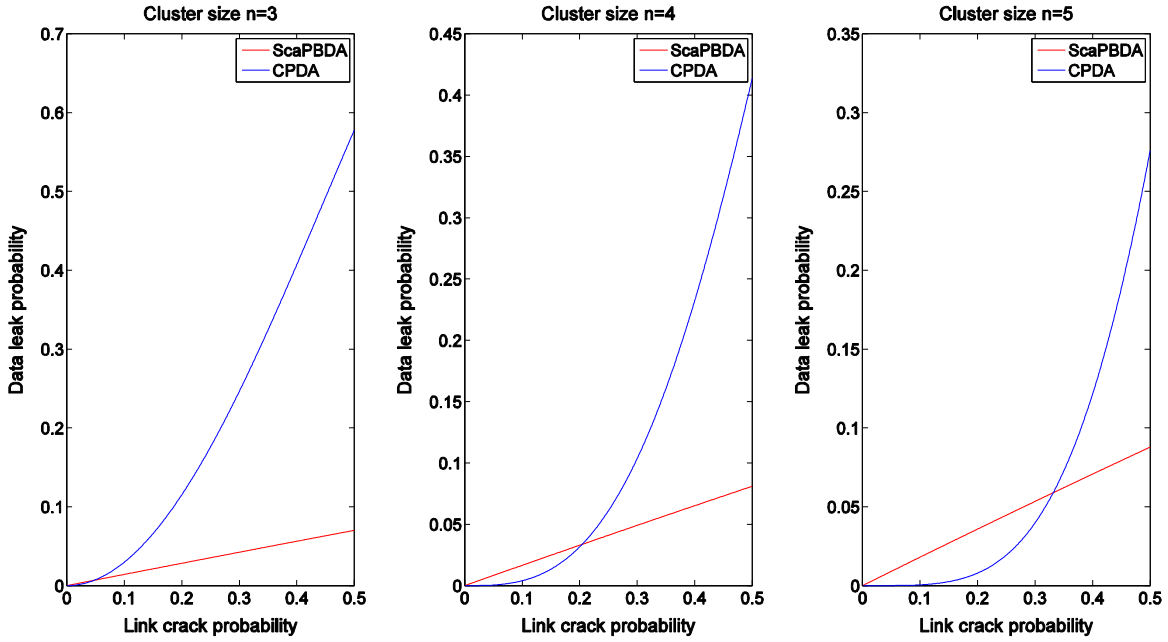

Fig. 6. Privacy performance.

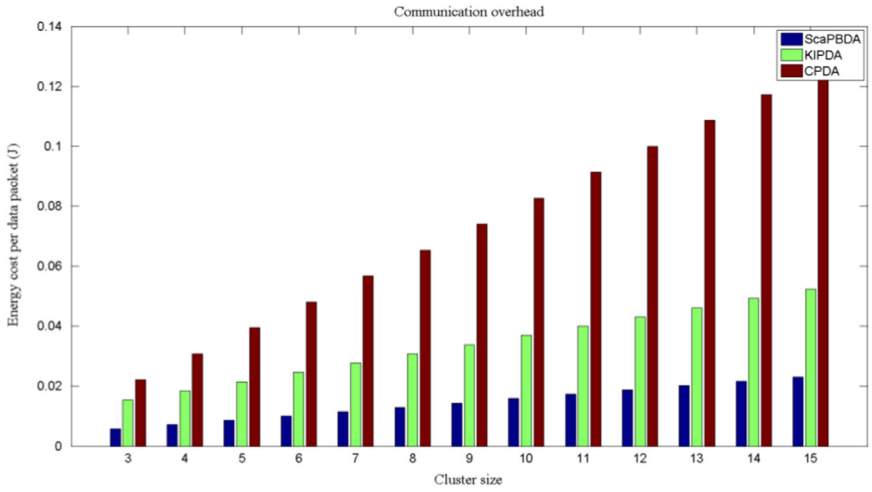

Fig. 7. Communication Overhead.

Given cluster size *CS*, the data customizing operations cause the same computation overhead in Sca-PBDA and KIPDA, and the total computation overheads of Sca-PBDA and KIPDA are almost the same and both are related to the packet length. Especially, the privacy-preserving configuration messages in both Sca-PBDA and KIPDA are assumed to be cryptographically and reliably transmitted. Therefore, they all have to be decrypted by the sensor nodes. However, the cryptographic CPDA is based on the frequent encryption and decryption operations by cluster members and the polynomials are employed to calculate the aggregation value, therefore its computational complexity is extremely high and its computation overhead is far more than those of Sca-PBDA and KIPDA. Once again, the communication overhead of large-scale WSNs is much larger than their computation overhead. Therefore, the proposed Sca-PBDA meets the application requirements of the computational complexity and scalability.

Lastly, the network lifetime is simulated to validate the efficacy of the proposed Sca-PBDA and compare it with related mechanisms. As shown in Fig. 8, the network lifetime is defined as the lifetime of the first energy-exhausted node, and the numerical trend of the alive nodes are clearly reflected. The balanced energy consumption of Sca-PBDA sufficiently extends the network lifetime, due to the proposed clustering methods with aCHs. However in KIPDA and CPDA, aggregating nodes with heavy energy consumption die much earlier, which severely affects the network connectivity and lifetime.

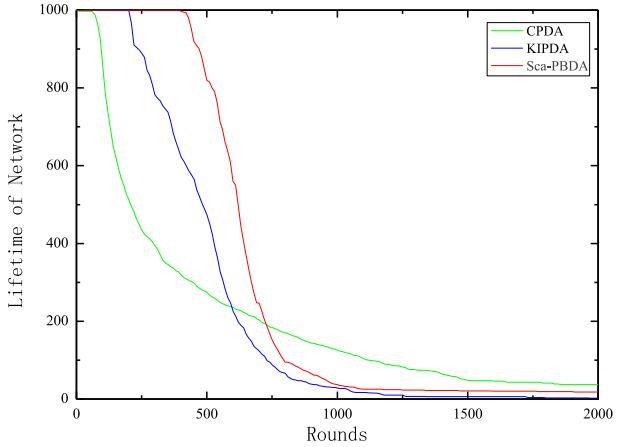

Fig. 8. Network lifetime.

7. Related works

The proposed Sca-PBDA is related to and inspired by many prior works on WSNs.

K-indistinguishable Privacy-preserving Data Aggregation (KIPDA) was proposed in [14] to achieve non-cryptographic data aggregation and transmission, which specially suits nonlinear aggregation functions. The proposed privacy-preserving data aggregation is more energy efficient than traditional hop-by-hop encryption methods. Cluster-based Private Data Aggregation (CPDA) was proposed in [15], which is based on the probabilistic election of cluster heads and the algebraic properties of polynomials. When calculating the in-cluster aggregation value, CPDA ensured that the seed-encrypted sensor data is unknowable to other cluster members. A hierarchical packet forwarding mechanism was proposed in [16] for energy harvesting WSNs, which provided inspiring ideas of gradient-aware unequal clustering methods. Ref. [17] introduced the Big data technology and its importance in the modern world and existing projects, which changed the concept of science into big science. Furthermore [18], presented the key issues of big data processing,introduced Map Reduce optimization strategies and the corresponding applications and gave the open issues, challenges and future research directions. Ref. [19] proposed a Big Data processing model, which involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. The concept of combining Hadoop and Storm was introduced in [20] for the gathering, storage and analysis of big sensor data generated by WSNs, and the sensor prototypes were also constructed in this paper. A novel framework integrating WSN with cloud computing model was proposed in [21] to offer reliable, easily accessible and extensible WSN services. A Dynamic Prime Number Based Security Verification (DPBSV) was proposed in [22] for big data stream processing, which greatly reduced security verification overhead and verification time. Besides, the security of the data was also enhanced by DPBSV. An aggregation strategy enabling the efficient use of energy and the handling of large data volumes was proposed in [23], and its basic concept is used to stream data from computer clusters or any network like structure. A scalable watermarking algorithm was proposed in [24] for authentications between personal mobile users and the media cloud, which not only achieved good security performance, but also enhanced media quality and reduced transmission overhead.

8. Conclusion

Targeting the privacy requirements of large-scale WSNs and focusing on the energy-efficient collection of big sensor data, Sca-PBDA is proposed in this paper. Firstly, a gradient-based equal network clustering method is proposed to reasonably determine the network topology, according to the estimated node energy consumption. With the proposed clustering method, the identical number of CH and CMs support the uniform privacy-preserving configuration and further inter-cluster data aggregation process, which meets the scalability requirements of the big sensor data collected by large-scale WSNs. Secondly, a scalable privacy-preserving data aggregation method is further designed to provide the simple privacy-preserving data configuration and scalable intra- and inter-cluster data aggregation. Especially, the effectively protected big sensor data is parallel aggregated at each CH, and relay CHs also perform aggregation operations on the received privacy-preserving aggregation data to further reduce the resource consumption. Lastly, aggregated results are recovered by the sink to complete the privacy-preserving big data aggregation. Simulation results validate the efficacy and scalability of Sca-PBDA and show that the big sensor data generated by large-scale WSNs is efficiently aggregated to reduce the network resource consumption and the sensor data privacy is effectively protected to meet the ever-growing application requirements. Most importantly, the proposed Sca-PBDA gives inspiring ideas of collecting and processing the big sensor data, where the energy-efficient parallel aggregating of big sensor data and the scalable privacy-preserving method for large-scale WSNs is of great reference value.

References

[1] D. Takaishi, H. Nishiyama, N. Kato, et al., Toward energy efficient big data gathering in densely distributed sensor networks, IEEE Trans. Emerg. Top. Comput. 2 (3) (2014) 388–397.

[2] S. Sagiroglu, D. Sinanc, Big data: a review. in: Proceddings of 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013, 42–47.

[3] I. Akyildiz, W. Su, Y. Sankarasubramaniam, et al., A survey on sensor networks, IEEE Commun. Mag. 40 (8) (2002) 102–114.

[4] I. Riazul, K. Daehan, K. Humaun, et al., The internet of things for health care: a comprehensive survey, IEEE Access 3 (1) (2015) 678–708.

[5] W. Richard, Winter Corporation Executive Report: Big Data: Business Opportunities, Requirements and Oracle's Approach. 2011, [online] Available: ⟨http://www.oracle.com/us/corporate/analystreports/infrastructure/winter-big-data-1438533.pdf⟩.

[6] P. Adrian, S. John, W. David, Security in wireless sensor networks, Commun. ACM 47 (6) (2004) 53–57.

[7] W. Yong, G. Attebury, B. Ramamurthy, A survey of security issues in wireless sensor networks, IEEE Commun. Surv. Tutor. 8 (2) (2006) 2–23.

[8] E. Fasolo, M. Rossi, J. Widmer, et al., In-network aggregation techniques for wireless sensor networks: a survey, IEEE Wirel. Commun. 14 (2) (2007) 70–87.

[9] R. Rajagopalan, P. Varshney, Data-aggregation techniques in sensor networks: a survey, IEEE Commun. Surv. Tutor. 8 (4) (2006) 48–63.

[10] O. Suat, X. Yang, Secure data aggregation in wireless sensor networks: a comprehensive overview, Comput. Netw. 53 (12) (2009) 2022–2037.

[11] Z. Zufan, Y. Yinxue, Y. Jing, Energy efficiency based on joint mobile node grouping and data packet fragmentation in short-range communication system, Int. J. Commun. Syst. 27 (4) (2014) 534–550.

[12] L. Yun, Z. Zhenghua, W. Chonggang, et al., Blind cooperative communications for multihop ad hoc wireless networks, IEEE Trans. Veh. Technol. 62 (7) (2013) 3110–3122.

[13] L. Changqing, L. Yang, L. Pan, et al., A holistic energy optimization framework for cloud-assisted mobile computing, IEEE Wirel. Commun. 22 (3) (2015) 118–123.

[14] M. Groat, H. Wenbo, S. Forrest, KIPDA: K-indistinguishable Privacy-preserving Data Aggregation in Wireless Sensor Networks, INFOCOM IEEE, 2011, 2024–2032.

[15] H. Wenbo, L. Xue, N. Hoang, et al., PDA: Privacy-Preserving Data Aggregation in Wireless Sensor Networks, INFOCOM IEEE, 2007, 2045–2053.

[16] W. Dapeng, H. Jing, W. Honggang, et al., A hierarchical packet forwarding mechanism for energy harvesting wireless sensor networks, IEEE Commun. Mag. 53 (8) (2015) 92–98.

[17] K. Avita, W. Mohammad, R.H. Goudar, Big data: issues, challenges, tools and good practices, in: Proceddings of the Sixth International Conference on Contemporary Computing (IC3), 2013, 404–409.

[18] J. Changqing, L. Yu, Q. Wenming, et al., Big data processing in cloud computing environments, in: Proceddings of the 12th International Symposium on Pervasive Systems, Algorithms and Networks, 2012, 17–23.

[19] W. Xindong, Z. Xingquan, W. Gong-Qing, et al., Data mining with big data, IEEE Trans. Knowl. Data Eng. 26 (1) (2013) 97–107.

[20] L. Rios, J. Diguez, Big data infrastructure for analyzing data generated by wireless sensor networks, in: Proceddings of 2014 IEEE International Congress on Big Data, 2014, 816–823.

[21] S. Shah, F. Khan, W. Ali, et al., A new framework to integrate wireless sensor networks with cloud computing, in: Proceddings of 2013 IEEE Aerospace Conference, 2013, 1–6.

[22] D. Puthal, S. Nepal, R. Ranjan, et al., DPBSV – an efficient and secure scheme for big sensing data stream, in: Proceddings of 2015 IEEE Trustcom/BigDataSE/ISPA, 2015, 246–253.

[23] R. Bergelt, M. Vodel, W. Hardt, Energy efficient handling of big data in embedded, wireless sensor networks, in: Proceddings of 2014 IEEE Sensors Applications Symposium (SAS), 2014, 53–58.

[24] W. Honggang, W. Shaoen, C. Min, et al., Security protection between users and the mobile media cloud, IEEE Commun. Mag. 52 (3) (2014) 73–79.