Research Article

Building Representative-Based Data Aggregation Tree in Wireless Sensor Networks

Yanfei Zheng, Kefei Chen, and Weidong Qiu

1 School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China
2 Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China

Correspondence should be addressed to Weidong Qiu, wdqiu@sjtu.edu.cn

Received 2 December 2009; Accepted 25 January 2010

Academic Editor: Ming Li

Copyright © 2010 Yanfei Zheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data aggregation is an essential operation to reduce energy consumption in large-scale wireless sensor networks (WSNs). A compromised node may forge an aggregation result and mislead base station into trusting a false reading. Efficient and secure aggregation scheme is critical in WSN applications due to the stringent resource constraints. In this paper, we propose a method to build up the representative-based aggregation tree in the WSNs such that the sensing data are aggregated along the route from the leaf cell to the root of the tree. In the cinema of large-scale and high-density sensor nodes, representative-based aggregation tree can reduce the data transmission overhead greatly by directed aggregation and cell-by-cell communications. It also provides security services including the integrity, freshness, and authentication, via detection mechanism in the cells.

1. Introduction

Wireless sensor networks (WSNs) have been used in many promising applications such as habitat monitoring, battlefield surveillance and target tracking. A larger number of tiny sensors collect measurement data and send them to processing center, which is usually called base station or sink node. However, the communication between sensors and processing center relies on multihop short range radio. As we know, WSNs also suffer from limited energy lifetime, slow computation, small memory, and limited communication capability. Obviously, the data aggregation can greatly reduce the communication consumption by eliminating redundant data in WSNs. It is known that aggregated traffic, modeled as fractal time series with complex characteristic [1, 2], has active applications in network security problems [3]. The study on aggregated traffic is significant in the wireless sensor network.
On the other hand, the sensors even the aggregators are vulnerable to attacks especially if they are not equipped with tamper-resistant hardware. When a sensor or an aggregator is compromised, it is easy for the adversary to inject bogus data into WSNs and change the aggregation results. Some methods have been proposed to solve the problem above. The works \cite{4, 5} use homomorphic encryption function to secure the aggregated data. The work in \cite{6} proposes secure aggregation tree without persistent cryptographs operations to detect and prevent cheating. The works \cite{7, 8} design the cheating detection mechanisms to guarantee the validation of aggregation data being sent in the WSN. The work in \cite{9–11} depends on the statistical feature of specific aggregation function. However, it is difficult to extend most of them to the large and high-density WSN due to the complexity of the operations or the structures.

In this paper, we propose a method to build the representative-based aggregation tree in the WSN on the basis of the work in \cite{8} such that the sensing data are aggregated along the route from the leaf cells to the root of the tree. In our scheme, the tree is not built directly on the sensors, but on the nonoverlapping cells which are divided with equal size in the target terrain. A representative sensor in each cell acts in name of the whole cell, including forwarding and aggregation of the sensing data in its cell and the receiving data from the neighbor cells. In the cinema of large-scale and high-density sensor nodes, our scheme cuts down the data transmission overhead from three aspects. The first is that the primary aggregation should be conducted in the cell, based on the observation that the measurement data in one small cell are almost identical. The second is that the aggregation operation in one large-scale network should be directed to avoid the dynamic change of the aggregation topology. The third is using cell-by-cell communication instead of hop by hop communication to reduce the density of communication and the complexity of the aggregation topology in the network. Also, each node in our scheme has a monitoring mechanism similar to the Watchdog that is proposed by Marti et al. \cite{12} in order to achieve cheating detection. The proposed scheme provides security services including the integrity, freshness, and authentication, via detection mechanism in the cells.

The rest of this paper is organized as follows. Section 2 presents our network model and notations in this paper. Section 3 gives the details of our scheme. Section 4 states the security and communications volume analysis. And section 5 concludes this paper.

2. Network Model and Notations

2.1. Network Model

We assume that the dimensions of the large deployment area are known in advance and the sensor nodes are uniformly distributed in this area. A grid structure is used to divide the target terrain into small nonoverlapping cells of equal areas as Figure 1 shows \cite{8}.

We assume that each node is aware of the dimension and the location of the cell to which it belongs. It is a reasonable assumption since the sensors with locating system are supported by most of current manufactories. It could also be deduced that the sensor node can judge which cells are its neighbor cells.

In our model, each cell has a cell representative which is selected based on its reputation, remained power, and so forth. A monitoring mechanism similar to Watchdog \cite{12} is set up in each node in order to monitor the cell representative operations. We donate
that two cells are the *neighbor cells* which have the adjacent edge. Due to the broadcast feature of radio channels, the messages in this sensor network are propagated along the neighbor cells—the cells with border line as Figure 2 shows. The dimension of each cell is small enough to allow the radio range of each node to cover its *neighbor cells*. The messages from non-neighbor cells, even if being detecting, will be ignored automatically by the receiving cells.

The base station is responsible for broadcasting the initial query to the monitoring area, processing received answers for these queries, and deriving meaningful information that reflects the events in the target field [8]. In order to simplify the illustration, the base station is located in the vertex of the target terrain even if base station is not there geographic. This assumption is reasonable since we can get one or more above graphs by rotating and dividing the coordinates if base station is in the boundary or middle of network.
2.2. Notations

The following notations are used throughout the paper.

- BS: The base station
- $x, y$: Sensor node $x$ and sensor node $y$, respectively
- $K_1, K_2$: Two network wide shared keys preloaded to each sensor node
- $n$: Total number of nodes in the target terrain
- $m$: Total number of cells in the target terrain
- $R_x, R_y$: Reading data from $x$ and $y$, respectively
- $C_i$: The $i$th cell
- $C_i^{\text{read}}$: Reported data from the cell $C_i$
- $C_i^{\text{rep}}$: The representative of the $i$th cell
- $T$: The number of nodes in each cell
- $t$: The minimal number of nodes in one cell requiring to revoke a new $C_i^{\text{read}}$
- $F$: An aggregated function
- $Q_n$: The $n$th query from BS
- $K_{C_i}$: Local cell key for the $i$th cell
- $K_{BS}$: Local cell key for the cell where BS locates
- $K_{C_i}^{C_j}$: Intercell key shared between the $i$th and the $j$th cell
- $\text{MAC}_{K_{C_i}}$: Message authentication code computed by using $K_{C_i}$
- $\text{MAC}_{K_{C_i}^{C_j}}$: Message authentication code computed by using $K_{C_j}^{C_i}$
- $\text{AD}_{C_i}$: Aggregation data from the cell $C_i$
- hop_count: The count of the cells on the route to the base station.

3. The Proposed Scheme

3.1. Main Idea

In a large-scale target terrain, the data aggregation may occur in any corner of network. The aggregating operation is also graduated up to the quietist. Obviously, the directed forwarding and aggregation along a steady skeleton have more advantages considering eliminating the duplicated sensing data and reducing energy consumption. Moreover, building such a steady skeleton in a hop-by-hop manner may not be a good choice in the situation of large-scale and high-density sensor nodes, which would lead to a deep and complex structure of the aggregation skeleton.

In our scheme, the queries from base station are spread along the cell by cell route. We build up an aggregation tree based on the cell. The aggregation data could be directionally delivered to the destination along nonoverlapped cells to avoid duplicated aggregation. The aggregation operations are conducted on each intermediate cell in the tree if necessary. A representative sensor in each cell acts in name of the whole cell, including forwarding and
aggregation of the sensing data in its cell and the receiving data from the neighbor cells. Other sensor nodes in the same cell monitor the behavior of their representative by listening to the communication between their representative and the representatives of the neighbor cells.

As the cheating detection mechanism is not the emphasis of our discussion, we build up the tree on the basis of the work in [8], whereas other monitoring mechanisms could also be used in our scheme. In this cheating detection mechanism, each node monitors the behavior of other nodes within the same cell and then calculates the reputation value for them based on their participation in some cell operations such as sensing, forwarding, and aggregating. If the current representative is detected to be compromised, the revocation mechanism will be started to generate the new representative.

3.2. Bootstrap

The bootstrap phase occurs in a short duration of time immediately after the network deployment. It is short enough to assume that no attacks are possible during this phase [8]. In this phase the local cell keys and intercell keys shared between two neighbor cells are established. Many works have been done on this kind of topics, such as [13–15]. We adopt the key distribution scheme stated in [8], which uses similar way as Ren et al. [16].

In this phase, each sensor node in the cell \( C_i \) computes the local cell key \( K_{C_i} \) which is used to authenticate any communication in the cell \( C_i \) by the following format:

\[
K_{C_i} = H(K_1 \| C_i),
\]

where \( \| \) represents bit string concatenation and \( K_1 \) is the preloaded key.

After that, each node in the cell \( C_i \) computes the intercell key \( K_{C_j}^{C_i} \) which is used to authenticate the communication between \( C_i \) and its neighbor cell \( C_j \) by the following format:

\[
K_{C_j}^{C_i} = H(K_2 \| C_i \| C_j),
\]

where \( K_2 \) is the other preloaded key.

At the end of this phase, each sensor node deletes \( K_1 \) and \( K_2 \) to prevent the adversary from getting access and sets its initial hop count value to infinity.

3.3. Cheating Detection Mechanism

To enhance the accuracy of the aggregated data without trimming the abnormal and bogus reading, the cheating detection mechanism based on the reputation proposed in [8] is introduced into our scheme. We briefly illustrate it for the completeness of the paper.

Since the local and intercell keys have been set up in the network after the bootstrap phase, the behavior of each node is under the detection of all the nodes in the same cell, including the cell representative. As soon as the reads of the cell departure the judgments of \( t \) nodes, the representative is responsible for computing the new cell reading. Each node establishes a reputation table to record the amount of positive and negative rating of every behavior of the other nodes in the cell. As soon as the reputation of the representative falls below a certain threshold, the revocation mechanism is triggered to generate a new representative based on the reputation records.
3.4. Building Representative-Based Aggregation Tree (RAT)

Before building RAT, we introduce the packet formats in the whole working phase. The packets have the following two formats:

\[ \{C_{rep}^i, C_{rep}^{parent}, [hop\_counts], flag, payload\}, \tag{3.3} \]

\[ \{C_{rep}^i, C_{rep}^j, Q_n, payload\}, \tag{3.4} \]

where \(C_{rep}^i\) is the representative of the sending cell \(C_i\), \(C_{rep}^j\) is the representative of the receiving cell \(C_j\), and \(C_{rep}^{parent}\) is the representative of parent cell of \(C_i\) in the tree.

Now, we propose a distributed algorithm to build RAT along the route of neighbor cells. The distributed algorithm builds the tree from the base station and includes the following steps.

**Step 1 (Invitation).** The base station locally broadcasts an invitation message to all of its neighbor cells, indicating that they should be its children. Since the base station is the root of the tree, it has no parent and its hop count is zero. The invitation message from the base station is described as the following format:

\[ \{BS, NON, 0, Invitation, payload\}, \tag{3.5} \]

where \(payload = MAC_{K_{BS}}(BS || Invitation)\).

A node \(C_{rep}^i\) in cell \(C_i\) who has joined the tree locally broadcasts the following invitation message:

\[ \{C_{rep}^i, C_{rep}^{parent}, hop\_counts, Invitation, payload\}, \tag{3.6} \]

where \(payload = MAC_{K_{C_i}}(C_{rep}^i || C_{rep}^{parent} || hop\_counts)\).

**Step 2 (Join).** Once the node \(C_{rep}^i\) in the neighbor Cell \(C_i\) receives an invitation message, if this cell has not joined the aggregation tree, the node \(C_{rep}^i\) records the sender of the invitation message as its parent node and updates its hop count value as one plus the hop count value in the received invitation message from its parent.

The node \(C_{rep}^i\) joins the tree by sending its parent the following join message:

\[ \{C_{rep}^i, C_{rep}^{parent}, “Join”, payload\}, \tag{3.7} \]

where \(payload = MAC_{K_{parent}}(C_{rep}^i || C_{rep}^{parent} || Join)\).

It is possible for a node to receive more than one invitation message. The node just takes the first invitation message as the active invitation due to the first invitation message would have the minimal hop count value normally. Once a node joins the tree, the later received invitation will be recorded for future use if its hop counts value is not bigger than the node’s current hop counts value.
The parent node will record its children by collecting the join messages. A cell is a leaf cell if it does not receive any join messages from any cell which announces to be its child.

**Step 3 (Iteration).** Repeat Steps 1 and 2 until all cells have joined the tree. The iteration process of invitation and join can be illustrated by Figure 3, where dot arrow line presents invitation message and arrow line presents join message.

### 3.5. Data Aggregation

An aggregation process begins when BS (the root of RAT) locally broadcasts a query $Q_0$ to its children by the following message:

$$ \{ \text{BS}, \text{all}_\text{children}, Q_0, \text{payload} \}, \quad (3.8) $$

where payload $= \text{MAC}_{K_{\text{BS}}}(\text{BS}||\text{all}_\text{children}||Q_0)$.

The data aggregation process includes the following two phases.

**Phase 1 (Query spread).** In the process of query spread, the intermediate cell propagates the query $Q_n$ to its children by the following message:

$$ \{ \text{C}_i^{\text{rep}}, \text{all}_\text{children}, Q_n, \text{payload} \}, \quad (3.9) $$

where payload $= \text{MAC}_{K_{\text{C}}}({C}_i^{\text{rep}}||\text{all}_\text{children}||Q_n)$.

The representative $C_i^{\text{rep}}$ in a leaf cell $C_i$ can propagate $Q_n$ to a node $x$ in its cell using the similar message format, if it randomly selects the reading $R_x$ as the reported data. Alternatively, it may take itself as the sensing node for simplify.
Phase 2 (Data aggregation). When a leaf cell \(C_i\) receives the query \(Q_n\), \(C^\text{rep}_i\) prepares \(C^\text{read}_i\) of some physical phenomena as queried and sends it back to its parent by the following message:

\[
\{C^\text{rep}_i, C^\text{rep}_\text{parent}, Q_n, \text{payload}\},
\]

where \(\text{payload} = C^\text{read}_i \parallel \text{MAC}_{K_{C^\text{rep}_\text{parent}}} (C^\text{rep}_i \parallel Q_n \parallel C^\text{read}_i)\).

If a node \(x\) is selected to report the sensing data in the last phase, it should report its reading to \(C^\text{rep}_i\) in advance by the following message:

\[
\{x, C^\text{rep}_i, Q_n, \text{payload}\},
\]

where \(\text{payload} = Rx \parallel \text{MAC}_{K_{C_i}} (x \parallel Q_n \parallel Rx)\).

When an intermediate cell receives the messages from its leaf, it verifies the MAC for the received data. If it does not match the received data, the reading from \(C^\text{rep}_i\) is ignored. Otherwise, \(C^\text{rep}_i\) considers the received data as an input to the aggregation function. After the \(C^\text{rep}_i\) receives all the readings or data from its children and the reading of the cell \(C_i\), it computes the result of the aggregation function as its report data and sends it to its parent by the following message:

\[
\{C^\text{rep}_i, C^\text{rep}_\text{parent}, Q_n, \text{payload}\}
\]

where \(\text{payload} = AD_{C_i} \parallel \text{MAC}_{K_{C^\text{parent}}} (C^\text{rep}_i \parallel Q_n \parallel AD_{C_i})\).

The data aggregation process also can be illustrated by Figure 3, where dot arrow line presents query message and arrow line presents aggregation message.

4. Discussions

4.1. Security Services Provided

The security requirements of data integrity, freshness, and authentication are achieved for the aggregation data in our scheme, since nodes share interkeys with the neighbor cells. As to the query message and the communication within the cell, the nodes in the same cell of the sender can authenticate it instead of the receiver, since nodes share local cell key in each cell. In fact, each monitoring node in the same cell can select some query messages randomly to low the energy consumption and prolong the lifetime of the cell. As the adversary is strong or the application is critical, the confidentiality could be achieved by encrypting the sensing data or aggregation data could be encrypted using inter/local keys.

4.2. Energy Cost Analysis of Data Transmission

Since the cell in RAT only communicates with the neighbor cells, the transmission distance is a constant value for each communication and the energy cost on data transmission is
mainly decided by the amount of data transmitted. So, we will discuss the data transmission volume of one response to a query in the RAT. For one response to query, two times of data transmissions are required in one cell. One is that the sensor node reports the sensing data to the representative of the cell. The other is that the representative of the cell reports the aggregation data up to its parent in the tree.

For the data aggregation function with fixed output size, such as min/max, the energy cost on data transmission with any aggregation tree is $2(m - 1)(b + c)$ where $m$ is the number of total cells, if we assume that $b$ is the size of measurement data and $c$ is the size of the subordinate in the message transmitted.

For some aggregation functions, the size of the return value is not fixed. It is a function of the total size of input data. We assume that such aggregation functions have fixed compression ratio of $\gamma$, where $0 < \gamma < 1$. As soon as sensing data pass by any one cell in the tree, they would be compressed once. Obviously, the total volume of data transmission in the tree depends on the structure of the aggregation tree. Assuming that each broadcast of the invitation message requires the same time, the first received invitation message should come along with the shortest path from the base station. The messages are only propagated along the neighbor cells that have the adjacent edge. Therefore the parent of one cell must be the upper neighbor or the left neighbor of its. Hence, the aggregation tree we build is an optimal tree with minimal communication cost.

Without losing generality, we still assume that $b$ is the size of measurement data and $c$ is the size of subordinate in the message. For a given aggregation tree, if denoting the maximal layer of the aggregation tree as $L$, total number of cell nodes as $m$, and the number of layer-$i$ node as $l_i$, the total transmission cost of an aggregation tree is given by

$$C = 2(m - 1)c + b \sum_{i=1}^{L} \sum_{j=1}^{l_i} \gamma^{j-1} l_i,$$

where $l_1 + l_2 + \cdots + l_L + 1 = m$.

Compared with the aggregation tree built on the hop-by-hop nodes in [6], the scale of RAT and transmission data are reduced to $1/T$ at least, in which $T$ is the number of nodes in one cell. The rate $1/T$ will be reached, if the hop-by-hop aggregation tree is also optimal and the output size of the aggregation function is fixed. Otherwise, we can get better result. In one word, our scheme shows a good property while facing the network of the large scale and high density.

The present results discussed above are assumed to be time invarying within a given interval of time, say $[0, I]$, if we select the starting point of time as 0. In this case, we take the parameter $I$ as the time scale at which the present algorithm is valid.

Note that a payload has complicated dynamics. The dynamics of traffic payload is strongly related to the selected time scale as can be seen from [17–23] and references therein. In general, traffic is nonlinear with fractal properties. In addition, it is nonstationary at small time scaling and stationary at large time scaling [24]. We now, taking (3.12) as an example, express the payload by

$$\text{payload}(m) = AD_{C_1} || \text{MAC}_{K_{C_{i_1}}}^{r_{\text{panal}}}(C_{i_1}^{r_{\text{rep}}} || Q_{i_1} || AD_{C_1})(m),$$

(4.2)
where $m$ ($m = 1, 2, \ldots$) is the index of the time interval. In this way, the previously discussed results should be all time varying with the index $m$, opening an attractive issue in the field. This point of view may be more agreement with the situation of real computer networks that are complex and dynamical in nature [25–31].

In future, we shall work on the statistics of the present algorithm from a view of nonlinear time series, which is challenging.

5. Conclusions

In this paper we have proposed a method of establishing the representative-based aggregation tree in WSN, where the network is divided into equal and nonoverlapping cells. In the cinema of large-scale and high-density sensor nodes, representative-based aggregation tree can reduce the data transmission overhead greatly by directed and cell-by-cell aggregating and forwarding. We have given the quantitative analysis of data transmission in the representative-base aggregation tree. At the same time, the monitoring mechanism in the cells prevents the injection of bogus information and forged aggregation values. In the future work, the problems which should be studied further are how to synthetically analyze the aggregated traffic in WSN from the aspect of fractal time series, including the traffic in RAT, to make further view of their characteristic of dynamics and nonlinearity.

Acknowledgments

This work was supported by the National High Technology Research and Development Program (863) of China under Grant no. 2009AA01Z418 and National Natural Science Foundation of China under Grant no. 60573125, no. 60873264 and no. 60903188.

References


