Two phase secure data collection technique for wireless sensor networks

Gousia Thahniyath1,2, Priti Mishra1, Sundar Raj Moorthy3
1Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Visvesvaraya Technological University, Karnataka, India
2Department of Computer Science and Engineering, Dayananda Sagar University, Karnataka, India
3Department of Biomedical Engineering, Rajiv Gandhi Institute of Technology, Visvesvaraya Technological University, Karnataka, India

ABSTRACT
Wireless sensor networks (WSNs) are the sensors that are dispersed in a different location that can sense the accumulated data in real-time and send it to the central location for the process of data aggregation. During the transfer of the information using the data nodes from the WSNs to the central location, there may be chances that the data node could be compromised by sensor failure or by an attack from a malicious user. To overcome this problem, we propose a two-phase secure data collection (TPSDC) technique for wireless sensor networks which provides confidentiality and integrity for the data nodes that are being sent to the central location and also during the data aggregation in the central location. Various existing methods have been proposed to secure the data when sent from the WSNs to the internet of things (IoT) devices but they lack to provide both confidentiality and integrity at the same time. Hence our model provides both integrity and confidentiality by providing security to the data nodes. Experimental results show that our model TPSDC performs better in terms of misclassification rate, detection rate, throughput, network lifetime analysis of the node, and communication overhead of the node when compared with the existing methods.

Keywords: Authentication, Consensus, Data aggregation, Integrity, Reputation system

This is an open access article under the CC BY-SA license.

1. INTRODUCTION
Wireless sensor networks (WSNs) are networks that are usually dispersed in many locations which can sense and collect data from various locations and forward it to a central location. WSNs are similar to ad-hoc networks because they rely on a wireless connection and a spontaneous formation of networks so that the sensed data using various sensors can be transported wirelessly to the central location. An internet of things (IoT) WSNs is usually used to collect the data and record the physical condition of the current environment where the WSNs have been placed and pass that information to an internet-based location. These WSNs can measure various kinds of data depending on the type of sensors. As to send the sensed data from the WSNs to the central location, the WSNs use various nodes to transfer the data from their location to the central location. During the transfer of the data, the sensor consumes memory, energy, computational speed, and communication bandwidth. As the transfer of data has to be done among the different nodes, there are some

Journal homepage: http://ijeecs.iaescore.com
chances that the data might get attacked by some malicious users and they could alter the data nodes sent by the sensor. To resolve this issue the data packets or the data nodes which are sent to the central location have to be secured with some security protocol that provides both confidentiality and integrity. In this method, some of the nodes which don’t have a proper trust between the central location get attacked. To overcome all these problems our model proposes a two-phase secure data collection (TPSDC) technique for wireless sensor networks which uses the sensors feedback information to secure the data nodes and also detects any unsecured data nodes and removes them to maintain confidentiality and integrity of the user. In Figure 1 it is shown how the data is aggregated using the WSNs and is forwarded to the central location (base station). WSNs connect each other to transfer the data from one sensor to another to reach the central location.

In the paper, Senturk et al. [1], an explanation of how the connectivity of mobile sensor networks can be restored by moving some nodes to the destination is done. Two existing methods considered fail to restore the connection between the nodes. The existing methods have been restructured so that they can determine the trajectory movement depending on the path planning algorithm. In the paper, Wang et al. [2], a method, hybrid recovery strategy based on random terrain in wireless sensor network utilizes the quantitative limits of the relay gadgets and realistic terrain influences for the restoration of the connection between the sender and the receiver. This method also reduces the cost of the energy required for data aggregation and collection. In this method, the approximation and complexity ratio have been discussed for their model. In the paper, Mi et al. [3], they have proposed a method, obstacle-avoiding connectivity restoration strategy, to resolve the problem of the failure of some sensors in the mobile robotic sensor networks. This method uses the backup selection algorithm for determining which sensor is currently being used and assigns a backup sensor next to it so that even when a failure occurs the backup sensor can extract the data. The selected sensor then avoids any obstacles using a gyro-sensor controller which restores the connection between the sender and receiver. In the paper, Yu et al. [4], an algorithm for the data aggregation has been proposed which utilizes the bitwise value of the XOR and provides a privacy-preserving min (i.e., minimal), percentile communication, and k-th min. This algorithm confirms whether the user data value is correct or not and also helps in the detection of whether the users are sending the non-repetition values such that it increases the accuracy in the data aggregation.

In the paper, Chen et al. [5], they have proposed two solutions for the problems in the data aggregation of the smart grids. They have proposed a method, data membership group-based multiple-data aggregation, which first divides the smart grid meters into different groups so that these groups can generate an encrypted key for their data and then dynamic leave and join along with the meter replacement methods are used for the data aggregation. In the paper, Yan et al. [6], for the fog nodes having untrusted servers, a data aggregation method for Fog-Assisted Mobile Crowd Sensing has been proposed for sharing the data among the users. The method preserves the privacy of the user’s data and the aggregated data results. This method provides reliability, privacy, and a secure communication metric for the Fog-assisted mobile crowdsensing. In the paper, Zhang et al. [7], a method for data aggregation has been proposed which uses the deep learning methods and compressed sensing capabilities to minimize the overall data which is transmitted to the IoT networks. In this method, the deep compressed sensing network has been utilized to attain a high accuracy reconstruction of the network using a measurement matrix. In the paper, B. Yin and X. Wei [8], to reduce the cost for the complex queries in the aggregation tree, a method has been proposed. In this method, the aggregation gain has been first formalized using the aggregation cost and data pruning power. After this, by exploiting the aggregation gain, the data has maximized high pruning power and small size is carefully chosen and moved to the data aggregation at the succeeding nodes. They have proposed a method that constructs the AT by connecting various sets of aggregation calculations attaining higher aggregation gains.

In the paper, Liu et al. [9], an efficient scheme using the blockchain data collection and deep-reinforcement-learning (DRL) methods has been proposed to develop a reliable and safe network for sharing and
exchanging data in mobile terminals. In this method, the DRL method is used to collect the maximum amount of data. This method uses the blockchain data collection method to provide reliability and security during the sharing and exchanging of data. The results have been evaluated in terms of reliability and security. The results show that this model performs better when compared with the existing database data sharing methods.

In the paper, Du et al. [10], for the security and data aggregation of the blockchain, a method, Spacechain, has been proposed. This method uses IoT devices for enabling the blockchain in the IoT. This method has proposed a scheme, three-dimensional greedy heaviest-observed sub-tree for the improvement of the network and to provide better security during the transmission of the data. In the paper, Yang, et al. [11], to resolve the problem of crowdsensing in the IoT devices, a scheme has been proposed that first identifies the three-way location which can be disclosed in the existing crowdsensing methods. To provide privacy among the users, they have proposed a method using the blockchain privacy methods which also helps to complete various tasks without any failure. To secure the transaction of the users they have proposed a private blockchain that can prevent the attack using re-identification. In the paper, Li et al. [12], an architecture using the blockchain, CrowdBC, a framework has been proposed which solves the user’s tasks using a crowd of employees without depending on any third-party application. This method helps them to maintain privacy and charge less amount of fees from the users. In the paper, Chen et al. [13], to provide more resources and security in IoT devices a data aggregation method has been proposed. In this method, a three-layer security framework using fog computing has been developed to provide integrity and confidentiality. For reducing the overall consumption of energy, an algorithm has been proposed which achieves a high convergence rate and an optimal value.

In the paper, Zhou et al. [14], an algorithm, energy-efficient, and privacy-preserving data aggregation have been proposed which consumes less energy and preserves privacy during the data aggregation. This algorithm slices the data acquired from each sensor to provide privacy for the data. In the paper, Chang, et al. [15], a method of consensus data aggregation that depends on Byzantine has been proposed. This method uses the threshold value of the data in the form of zero and one which helps in the aggregation of the data. This method reduces the consumption of energy and helps to forward the data to the consensus at a high speed. It has also solved the problem of fault tolerance and failure of nodes. In the paper, Banerjee et al. [16], have modified the low energy adaptive clustering hierarchy protocol to reduce the consumption of energy and to improve the performance of the network. In the paper, Yuwen et al. [17], they have presented two methods for data aggregation which preserve the privacy of the user’s data. In this method, they have sliced the data and encrypted using the advanced encryption standard (AES) encryption to secure the communication between the sensor device and the IoT devices. In the paper, Dou et al. [18], they have proposed an algorithm, secure and efficient privacy-preserving data aggregation algorithm (SECPDA), which provides privacy to the clustered data aggregation data. This algorithm selects the cluster head nodes and uses various slicing methods to provide privacy on the aggregated data. In the paper, Faris et al. [19], they have proposed an authentication method for the healthcare application, efficient and privacy-preserving data aggregation scheme with authentication for IoT-based healthcare applications, which verifies the nodes and detects which node needs to be processed for data aggregation. It provides security to the model using the homomorphic MAC protocol which in turn provides integrity to the data.

2. METHOD

In this section, the data aggregation for the wireless sensor network to provide security and integrity has been discussed. In this research, a two-phase framework to provide integrity for secure data aggregation in the wireless sensor network has been given. The two-phase framework is shown using Figure 2. In Figure 2, Phase-1 first collects all the feedback information taken by the wireless sensor network through the IoT devices, validates the feedback, and provides a trust-based communication metric by considering direct, indirect, and biased feedback. After Phase-1, in Phase-2 the aggregated data is secured using our improved consensus model which detects the insecure data packet and removes all the insecure data packets. In this research, the framework first develops a trust-based communication metric using the feedback information taken by the WSNs from the different sensor nodes. After the trust-based communication metric has been well-known, the secure aggregation of the data is executed. For the secure aggregation of the data, this model first detects all the insecure data packets and removes them to save energy efficiently.

2.1. System model

This work collects the sensory data collected through different workflow models as described in [20], [21]. All these workflow model requires a secure aggregation model [21]. Thus, this paper presents a secure aggregation model for clustering-based WSNs. The CH that performs aggregation cannot be trusted.
In this work, two different nodes exist such as malicious and genuine ones. Thus, we have two issues, first, degradation in the quality of data collected as malicious sensor nodes might induce information tampering. Second, sensor device confidentiality may be compromised through eavesdropping.

![Two-phase framework model for secure network to provide integrity in WSNs](image)

2.2. Consensus-based efficient and secure data aggregation

This paper presents the consensus-based efficient secure data aggregation scheme for WSNs, namely TPSDC. The aggregated data $z^l$ is computed using the information of $j^{th}$ sensor node within $l^{th}$ session instance $y_j^l$ and respective weights $x_j^l$ as (1).

$$z^l = \sum_{j=1}^{O} x_j^l y_j^l,$$

(1)

Every sensor node $w_j$ adds noisy data $\delta_j$ that follows Gaussian distribution to the actual data $y_j$ is functionally defined as (2):

$$\tilde{y}_j = y_j + \delta_j$$

(2)

where $\delta_j \sim O(0, \sigma^2)$. Therefore, the (1) is updated as (3):

$$\hat{z}^l = \sum_{j=1}^{O'} x_j^l \tilde{y}_j^l,$$

(3)

where $O'$ defines the trustable between $O$ communicated information. The noise introduced in (3) is done through random function $N(\cdot)$ to information $y_j$ and is functionally defined as (4).

$$\tilde{y}_j = N(y_j) = y_j + \delta_j$$

(4)

To increase security by preserving confidentially here an incentive parameter $\zeta$ is used as (5):

$$\zeta = |\bar{z} - \hat{z}|$$

(5)

where $\bar{z}$ defines the true mean with respect to the outcome $\hat{z}$. The lesser value of $\zeta$ indicates the higher security of the WSNs. The aggregated data is reliable and can be validated only when it has a good quality of data aggregation. Hence, to detect the dishonest nodes in the aggregated data a method has been given in this model to provide good quality of data aggregation. For this method, consider $J_0$ constraint which provides better efficient data during the data aggregation and $J_1$ constraint which provides non-efficient data during the data aggregation. Using these considerations, we can evaluate all the misclassified packets (nodes) which are trustable for the communication metric. This evaluation expression is given as (6).

$$R_h = R(J_1|J_0).$$

(6)

Furthermore, the rate of misclassification can be given as $R_m$ in which the non-trustable nodes are taken into consideration for the trustable ones which are given using (7).
\[ R_m = R(J_0|J_1). \] (7)

According to (7), the test static \( M \) is designed and given using (8).

\[ M = \| y_j^1 - \hat{y}_j^1 \|^2 \] (8)

The (8) gives the deviation between the two terms defined in (6) and (7). Consider an overall aggregated data which comprises all the sensed, noise, and additional data. This overall data can be expressed using (9).

\[ y_j = (y_1^j, y_2^j, \ldots, y_o^j) \] (9)

For this overall aggregated data given in (9), the test for the classified and misclassified nodes is designed as (10):

\[ M \leq J_0 (\theta) \] (10)

where \( \theta \) defines the trust factor of aggregated data. Hence if the aggregated data is trustable, the \( y_j^1 \) the number of sensor nodes is updated or else it will be removed. This can be denoted as (11):

\[ y_j^1 \leftarrow y_j^1 \] (11)

otherwise,

\[ y_j^1 \leftarrow y_j^{1-1} \] (12)

to evaluate the attack risk, consider an energy constraint \( E \) which consists of all the absolute nodes given as \( \mathbb{I}_1 \) and consider the dishonest nodes as \( \mathbb{I}_0 \). These considerations are expressed as (13).

\[ \mathbb{I}_1 > \mathbb{I}_0 > 0 \] (13)

Consider an attack probability parameter which is represented as \( r \) and the probable risk for the attack which is represented using \( S(\theta, r) \). Using these an equation can be formulated for the attack risk which is given using (14).

\[ S(\theta, r) = (\mathbb{I}_1(1 - R_h(\theta)) - ER_h(\theta))(1 - \sum_{j=1}^{o_n} r_j) + (\mathbb{I}_0(1 - R_m(\theta)) - ER_\hat{h}(\theta)) \sum_{j=1}^{o_n} r_j \] (14)

In the (13), by considering the clustering utility, \( v_c(\theta, R) \), it can be seen that \( v_c(\theta, R) = S(\theta, R) \). In the next sub-section, a secure metric using a trust-based model is given.

### 2.3. Secure metric for classifying malicious sensor nodes

In attaining better communication among sensor nodes, a secure communication metric \( F_o^u(x, y) \) is defined using [22], [23] as (15).

\[ F_o^u(x, y) = E_o^u(x, y) * \mathbb{I}_o^u(x, y). \] (15)

Using (11), it can be represented that the IoT device has established the trust between the nodes having a secure communication metric and integrity. The (14) also selects the IoT devices which have more trust between the nodes which comprises both the security and integrity [24].

In this section, the secure communication metric for the sensor nodes is calculated. Using (14), the trust between the nodes is identified. When the nodes are sent to the IoT devices using the wireless sensor network, it utilizes more energy above the IoT device in the cluster network having more trust between the nodes [25]. Hence, to balance the load between the cluster head, the energy consumed during the traffic between the nodes is calculated as (16):

\[ T^u(x, y) = T^u(x, y) + \sum_{p \in Z - \{x\}} F^u_o(x, p) * T^u(p, y) \] (16)
using (15), when the traffic between the nodes is calculated, the selection of the cluster head to transmit the
nodes to the IoT device is calculated as (17):

\[ \min \sum_{p \in \mathbb{Z} - \{x\}} T^u(x, p) \]  

(17)

after the selection of the cluster, the head has been evaluated, if any of the IoT devices don’t have any trust
value, then the probability of the IoT device is evaluated as (18):

\[ \mathcal{P}^u(x, y) = \begin{cases} \frac{\mathcal{F}^u(x,y)}{\sum_{p \in \mathbb{V}} \mathcal{F}^u(x,p)}, & \text{if } \sum_{p \in \mathbb{V}} \mathcal{F}^u(x,p) \neq 0, \\ \text{arbitrarily choose any sensor device, else}. \end{cases} \]  

(18)

for the selection of the IoT device, this model uses (17) which gives a high probability for the selection of the
IoT device. If the trust-based probability parameter is set to 0 or is equal to 0 then the IoT device is selected
at a random instance. Furthermore, both these two-phase frameworks provide security and integrity for the
model and the results show high performance during the data aggregation which is shown in the next section
of results and discussions.

3. RESULTS AND DISCUSSION

In this section, the two-phase secure data collection technique for wireless sensor networks has been
compared with the existing system. The model has been compared using the following considered terms:
misclassification rate, detection rate, throughput, network lifetime analysis of the node, and communication
overhead of the node. The experimentation results have been discussed to prove that our TPSDC is more
efficient in many ways when compared with the existing systems.

3.1. Rate of misclassification

In this section, the misclassification rate of the nodes varied with the malicious nodes has been
discussed. Figure 3 shows how our two-phase secure data collection model has a lower misclassification rate
when the malicious nodes increase. It can be seen from Figure 3 that the two-phase secure data collection has
a lower misclassification rate when compared with the existing system. As the malicious node increases the
misclassification rate also increases constantly. Our model attained 16%, 21%, 35%, and 55% for the
misclassification rate in 5%, 15%, 25%, and 35% of malicious nodes respectively whereas for the existing
system, it was 28%, 37%, 53%, and 69% for the misclassification rate in 5%, 15%, 25%, and 35% of
malicious nodes respectively.

![Figure 3. Misclassification rate varied with malicious nodes](image)

3.2. Detection rate

In this section, the detection rate of the nodes has been discussed varied with the percentage of
malicious nodes. Figure 4 shows how our two-phase secure data collection model has a higher attack
detection rate when the malicious nodes are fewer and gradually decreases as the malicious node increase.
From the results acquired which can be seen from Figure 4, it can be said that our model two-phase secure
data collection detects the attack more precisely when compared with the existing system. As the malicious
nodes increase the attack detection rate decreases in both the existing system and our two-phase secure data
collection model. Our model attained 84%, 79%, 65%, and 45% for the attack detection in 5%, 15%, 25%, and
and 35% of malicious nodes respectively whereas for the existing system 72%, 63%, 47%, and 31% for the attack detection in 5%, 15%, 25%, and 35% of malicious nodes respectively.

3.3. Throughput

In this section, the throughput of the nodes varied with the malicious nodes has been discussed. Figure 5 shows how our two-phase secure data collection model has higher throughput when the malicious nodes are less and gradually decreases as the malicious node increase. From the results acquired which can be seen from Figure 5, it can be said that our model two-phase secure data collection has higher throughput when compared with the existing system. As the malicious nodes increase the throughput decreases in both the existing system and our two-phase secure data collection model. The existing model attained a throughput of 0.612, 0.4788, 0.2725, and 0.1209 for the 10, 20, 30, and 40 nodes whereas our model two-phase secure data collection attained throughput of 0.714, 0.6004, 0.377, and 0.1755 for the 10, 20, 30, and 40 nodes respectively.

3.4. Network lifetime analysis

In this section, the network lifetime analysis of the nodes has been done to discuss the loss of connectivity. In Figure 6 it can be seen that our model two-phase secure data collection has less loss connectivity when the number of rounds increases. Our model provides more lifetime for the given node before disconnecting and is constant even when the number of sensor nodes increases. In the existing system, the network lifetime increases and suddenly decreases as the number of sensors nodes changes.

3.5. Communication overhead

In this section, the communication overhead of the nodes has been evaluated depending on the number of sensor nodes. In Figure 7 the communication overhead of the existing system and our two-phase secure data collection model has been shown. It can be seen that the communication overhead is constant and does increase gradually as the number of sensor nodes increases whereas in the existing model the communication overhead increases as the number of sensor nodes increases. From all the above sections it can be discussed that our model two-phase secure data collection performs better in terms of misclassification rate, detection rate, throughput, network lifetime analysis of the node, and communication overhead of the node when compared with the existing system.
4. CONCLUSION

Data collection in the wireless sensor networks is the main fundamental module. When the data is collected and aggregated, the data has to be secured, have confidentiality and integrity. Some of the existing models provide integrity by preserving the privacy of the user’s data but fail to provide confidentiality. Some of the models provide confidentiality but have failed to provide integrity both at the same time. Hence, our model two-phase secure data collection provides both at the same time. In our model, first, the feedback information from the WSNs has been extracted and authenticated. After that, the data nodes have been secured and forwarded to the central location. If the packets are unsecured, then our model detects the unsecured packets and removes them if required. The model has been evaluated using the following terms: misclassification rate, detection rate, throughput, network lifetime analysis of the node, and communication overhead of the node. Our model, TPSDC has performed better in all the above terms mentioned when compared with the existing system.

REFERENCES

BIOGRAPHIES OF AUTHORS

Gousia Thahniyath is Assistant Professor, Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore. Gousia Thahniyath is pursuing her Ph.D. from Visveswaraya Technological University. Her area of research is related to Wireless Sensor Networks. She has completed her M.E in Computer Science and Engineering from Bangalore University and B.E in Computer Science and Engineering from Gulbarga University. She has 14 years of teaching experience and has one year of industrial experience. She has published a paper in 7 International Journals and has presented papers in International and National conferences. She can be contacted at email: gous.tans@gmail.com.

Dr. Priti Mishra earned a bachelor's degree in information science and engineering from SRSIT College of Engineering in Bangalore, which is affiliated with VTU. MVJ College of Engineering, affiliated to VTU, Bangalore, and Maharaj Vinayak Global University, Jaipur, respectively, with a PG in Computer Science & Engineering and a Ph.D. in Computer Science, with a focus on network security. She's been a teacher for the past 16 years. In several technological organizations, she served in various roles such as Associate Professor and Professor. She is currently employed at Rajiv Gandhi Institute of Technology, Bangalore, as a Professor in the Department of Computer Science and Engineering. She can be contacted at email: sitizaiton@umk.edu.my.

Prof. Dr. Sundar Raj Moorthy, Professor & Head, Department of Biomedical Engineering, Rajiv Gandhi Institute of Technology, Bangalore, India. He has completed his BTech (2009) in Biomedical Engineering from Bharath University, India, MTech (2012) in Biomedical Engineering from Vellore Institute of Technology, India, and Ph.D., ECE/Biomedical Engineering (2017) from Bharath Institute of Higher Education & Research, India. He can be contacted at email: sundarmbe@gmail.com.