Ensuring Data Integrity by Anomaly Node Detection during Data Gathering in WSNs

Quazi Mamun, Rafiqul Islam, and Mohammed Kaosar

School of Computing and Mathematics, Charles Sturt University, NSW, Australia
{qmamun,mislam,mkaosar}@csu.edu.au

Abstract. This paper presents a model for ensuring data integrity using anomalous node identification in non-homogeneous wireless sensor networks (WSNs). We propose the anomaly detection technique while collecting data using mobile data collectors (MDCs), which detect the malicious activities before sending to the base station (BS). Our technique also protects the leader nodes (LNs) from malicious activities to ensure data integrity between the MDC and the LNs. The proposed approach learns the data characteristics from each sensor node and passes it to the MDC, where detection engine identifies the victim node and eventually alarm the LNs in order to keep the normal behaviour in the network. Our empirical evidence shows the effectiveness our approach.

Keywords: WSN, data integrity, mobile data collector, compromise, malicious, anomaly.

1 Introduction

The development of WSNs has attracted a lot of attentions due to the potentiality of broad applications in both military and civilian operations. Usually WSNs are deployed in unattended and often hostile environments such as military and homeland security operations [1, 2, 3, 6, 17, 18, 19, 22]. Recent advances in wireless sensor network research have shown that an attacker can exploit different mechanisms of sensor nodes spread malicious code through the whole network without physical contact [18, 19]. Therefore, it is imperative to adopt security mechanisms providing confidentiality, authentication, data integrity, and non-repudiation, among other security objectives, are vital to ensure accurate network operations.

A WSN may consist of hundreds or even thousands of sensor nodes. The sensor node consists of distributed autonomous devices using sensors to cooperatively monitor or collect sensing data at different locations. This renders it impractical to monitor and protect each individual node from a variety of malicious attacks. For instance, once a particular node is compromised, intruders can launch various malicious codes to launch attacks. They might spoof, alter or replay routing information to interrupt the network routing [1]. They may also launch the Sybil attack [2, 3], where a single node presents multiple identities to other nodes, or the

T. Zia et al. (Eds.): SecureComm 2013, LNICST 127, pp. 367–379, 2013.

identity replication attack, in which clones of a compromised node are put into multiple network places [3]. Moreover, adversaries may inject bogus data into the network to consume the scarce network resources [4, 5]. In addition, if the coordinators of the sensor networks are compromised, all members of the clusters become more vulnerable to different types of security attacks. This situation poses the demand for compromise-tolerant security design, especially for the coordinators.

A node can be captured by the intruder to find sensitive data or to compromise other nodes. In all of the WSN topologies, the sensor nodes send the sensed data to the coordinators. In a clustered based topology, the coordinators are called cluster heads, whereas in a chain oriented network the coordinators are called chain leaders. If the coordinator is compromised, these nodes can be used by the intruders to compromise other nodes. Thus coordinator compromise is a serious threat to wireless sensor networks deployed in unattended and hostile environments. To mitigate the impact of compromised nodes, we propose a model of compromise-tolerant security mechanism by adopting a detection engine within the MDC. This technique will enable to protect the BS as well as cluster coordinator and ensure the data integrity between MDC and BS.

The main contributions of this paper are three-folds:

- The design of deploying detection engine within the mobile data collectors to identify the malicious node of WSN cluster. As mobile data collection techniques are attracting attentions nowadays due to their energy-saving characteristics, the proposed idea is well fitted in this category of research.
- The proposed method prevents not only the members of a cluster or chain, but also the leader of the cluster/chain from malicious activities. Additionally, as the mobile data collectors are carrying messages to the BS, the BS can also be kept safe from the malicious activities.
- The proposed method reduces memory overhead. We adopt the mobility in collecting data by utilizing multiple mobile data collectors (MDCs) and enhanced the performance of data collection process by using the spatial division multiple access (SDMA) technique.

The rest of the paper is organized as follows. Section 2 presents the network architecture model of the proposed node anomaly detection technique. Section 3 describes details of our anomalous node detection method. In Section 4, we describe the experimental setup. Simulation results are presented in Section 5. Finally in Section 6, we draw the conclusion and describe the future work.

2 Network Architecture Model

The proposed anomaly node detection technique can be deployed over all hierarchical networks, either in cluster based or tree based or chain oriented network. In this paper, we consider the network topology is chain oriented topology. In a chain oriented sensor network, multiple chains can be constructed, where all the chains will be restricted to *Voronoi cell*s [22]. Furthermore, in these topological networks, mobile data collectors can be used to collect data from the deployed sensor nodes [8].

An overview of the architectural model is illustrated in Figure 1. The leader nodes are depicted using the blue coloured dots. All the sensor nodes deployed inside a

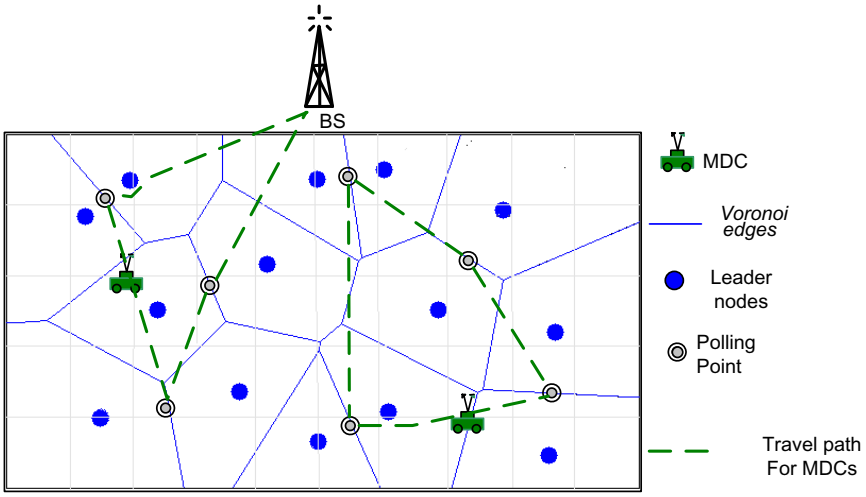


Fig. 1. The network architecture model for the proposed anomaly node detection technique

Voronoi cell send their data to the leader nodes. On the other hand, the mobile data collectors visits the polling points on a regular basis and collect data from the leader nodes. The data gathering scheme for large scaled wireless sensor networks can be extended by using multiple MDCs and the spatial division multiple access (SDMA) technique. This is described in details in [8]. For example, in Fig. 1, two MDCs travel within the network and collect data from the leaders. The two MDCs work at the same time, and when an MDC arrives at a polling point, leaders associated with this polling point are scheduled to communicate with the MDC. Two leaders in a compatible pair can upload data simultaneously in a time slot, while an isolated leader (i.e., a leader by itself or not in any compatible pair) sends data to the MDC separately.

The BS is usually situated outside the sensing field. Sending the data by the sensors to the remote BS may lead to non-uniform energy consumption among the sensors, because the sensor nodes (or leader nodes) that are responsible for sending data to the BS, need to cover long-range distances. As a result, they deplete energy much faster than other sensor nodes, and die quickly [9, 10, 11]. The consequence of this situation may result in partitioning the network and loss of robustness. However recent studies [12, 13, 14] have proposed sink mobility or collecting data using a mobile device as an efficient solution for data gathering problem. Employing mobile devices to collect data can reduce the effects of the hotspots problem, balance energy consumption among sensor nodes, and thereby prolong the network lifetime to a great extent [15, 16].

To solve the vital data gathering problem of large scaled WSNs, we adopt mobility in collecting data by utilizing multiple mobile data collectors (MDCs) and enhanced the performance of data collection process by using the spatial division multiple access (SDMA) technique [8]. In the proposed scheme, the sensing field is divided into several non-overlapping regions and for each of the regions, an MDC is assigned.

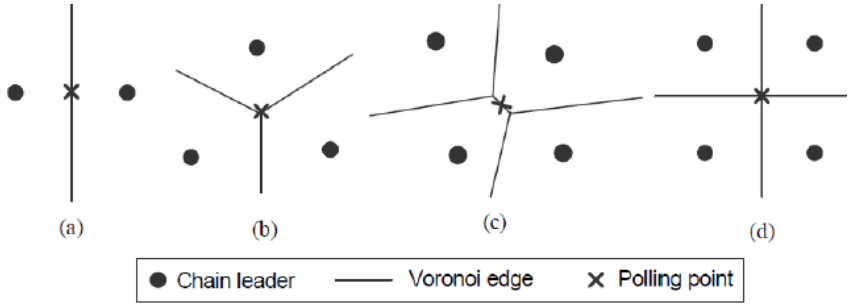


Fig. 2. Polling points are marked on the Voronoi edges

Each MDC takes the responsibility of gathering data from the leaders in the region while traversing in their transmission ranges. The traversal paths of the MDCs are determined using the *Voronoi* diagram constructed with respect to the leader nodes. We also consider exploiting the SDMA technique by equipping each MDC with two antennas. With the support of SDMA, two distinct compatible leader nodes in the same region can successfully make concurrent data uploading to their associated MDC. Intuitively, if each MDC can simultaneously communicate with two compatible leader nodes, the data uploading time in each region can be cut in half in the ideal case.

We further focus on the problem of minimizing DGS time among different regions. Besides this, the data gathering problem using multiple MDCs and the SDMA technique requires optimal solutions, discussed in [8, 13]. These optimization problems can be formulated using an Integer Linear Programming (ILP) approach. However, the complexity of an ILP solution is generally high, which is not suitable for a large scaled WSN [11]. Therefore, a heuristic region-division and traversing algorithm was used in [8] to provide a feasible solution to the problem. One of the common challenges of the WSN is the conservation of power, thus elongating the life span of a sensor node. A lot of research is being carried out towards 'energy efficiency' of WSN. In this paper an attempt has been made to secure the WSNs with the help of 'Cross layer' approach. This is an extended and enhanced version of [8].

In our proposed model, an MDC travels within each region and stops at some locations to collect data from the leader nodes. These positions are called polling points. To take full advantage of the SDMA technique, polling points should be equidistant from the associated leader nodes. Figure 2 shows some positions of polling points in four different cases. If there are only two leader nodes, the position of the polling point can be found at the intersection between the *Voronoi* edge and the line joining the two leader nodes (Fig. 2(a)). For more than two leader nodes, the

polling point can be found at the intersection of different *Voronoi* edges (Fig. 2(b-d)). Any two leader nodes associated with the same polling point are said to be compatible if an MDC arriving at this polling point can successfully decode the multiplexing signals concurrently transmitted from these two leader nodes. Detailed discussions on utilizing SDMA at physical layer for concurrent data uploading is provided in [8].

The following assumptions are made specific to our proposed DGS:

- It is assumed that MDCs have access to a continuous power supply. Usually the BS is equipped with the source of continuous power supply. Thus, when an MDC visits the base station, it can replace its battery.

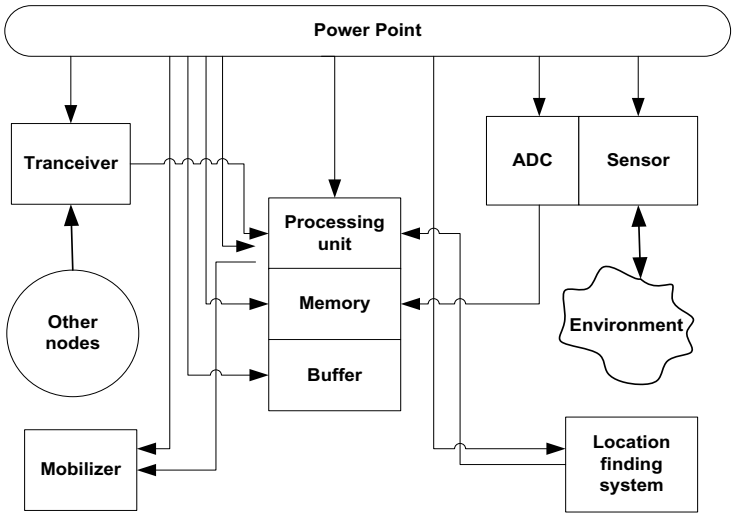


Fig. 3. Basic structure of a sensor node

- It is assumed that the MDCs are familiar with the target field. Location images of the target field can be stored in each MDC. Thus, an MDC is able to visit any point within the target field.
- It is also assumed that each MDC can forward the gathered data to one of the nearby MDCs when they are close enough, such that data can eventually be forwarded to the MDC that will visit the static data sink.

3 Anomaly Node Detection Method

In this section we present the scaffold of our anomaly detection technique. First we present the basic diagram of a sensor node which integrates hardware and software for sensing, data processing, and communications. They rely on wireless channels for transmitting data to and receiving data from other nodes. A sensor node is made up of a sensing unit, a processing unit and transceiver unit and a power unit, as illustrated in Figure34. They may also have additional application-dependent components such as a location finding system, power generator and mobilize. Sensors devices that can observe

or control physical parameters of the environment is converted to digital signals by the ADC, and then fed into the processing unit. The processing unit which is generally associated with a small storage unit, manages the procedures that make the sensor node collaborate with the other nodes to carry out the assigned sensing tasks. A transceiver unit connects the node to the network. Power units may be supported by power scavenging units such as solar cells. Most of the sensor network routing techniques and sensing tasks require knowledge of location with high accuracy. Thus, it is common that a sensor node has a location finding system. A mobilize may sometimes be needed to move sensor nodes when it is required to carry out the assigned tasks.

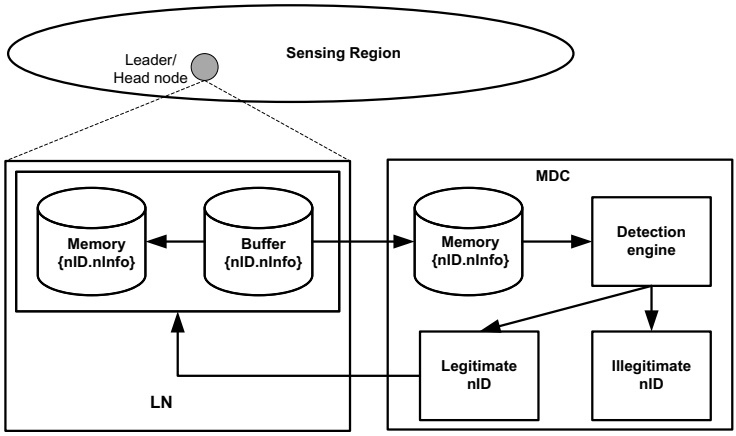


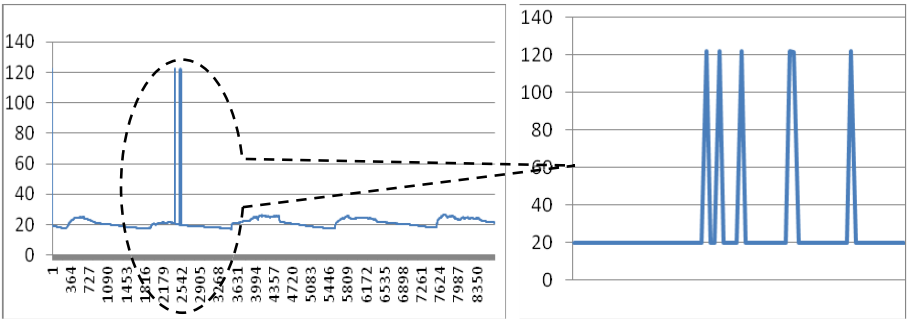
Fig. 4. Anomaly detection model

Figure 4 illustrates the overview of our proposed detection model. We propose two approaches for detecting the malicious node in our model. Firstly leader node, we called it LN, which will store the receiving data from all sensor nodes in the cluster into their buffer memory, as shown in Figure 4. Then the MDC node should use another buffer where the sensing data, from the LN, will be stored. In both cases every node ID (nID) will be considered as uniquely identifier of each node, so that after detection LN can identify the victim node. The detection engine will be deployed into the MDC node and the data will be passed to detection engine to identify the victim node. Finally the MDC will transfer the list of legitimate nID to the LN and then LN will transfer the data to its main memory based on the list of supplied nID. The MDC also send the information to BS based on the legitimate nID. If there is any malicious node identified, the MDC will immediately inform to the LN for taking protection measure.

4 Experimental Setup

The purpose of this experiment is to evaluate the effectiveness of anomaly detection of LN within WSN region. Our evaluation is based on a real-life dataset in which the modes or partitions in the data can be controlled.

We use a real-life dataset called the IBRL dataset in our evaluation [21]. The IBRL data set includes a log of about 2.3 million readings collected from 54 sensor nodes. The total log size is 150MB and the data were averages averaged over all time. The IBRL data is a publicly available set of sensor measurements gathered from a wireless sensor network deployed in the Intel Berkeley Research Laboratory [21]. In this data set the have used temperature and humidity data of 12 hour periods. In this period, as shown in Figure 5, one of the sensors started to report erroneous data or abnormal data. This can be seen as a dotted block in figure 5(a) and the elaboration in figure 5(b). Analysing the behaviour of the sensors showed that most had such behaviour toward the end of the experiment, but this particular sensor started its drift earlier than the others.



(a) (b)

Fig. 5. Abnormal data reading from sensors

To investigate the effect of a non-homogeneous environment, a synthetic dataset, from real dataset, with five disjoint clusters was built. Data from each sensor was gathered randomly according to the distribution (cluster) assigned to that sensor. The data is generated so that it has the same range as the IBRL dataset. Since each sensor recorded multiple readings of the same temperature and humidity, we can compress the data. Instead of keeping all (temperature, humidity) attributers, only unique attributes have been kept with their relative frequency of occurrence. Also, temperature and humidity values have been rounded up to whole numbers. With this method, the volume of the data has been reduced significantly by over an order of magnitude.

Detection Engine: The first step of our detection engine is to select the parameters to monitor and group them in a pattern vector [x1] $x^{\mu} \in \Re$, $\mu=1,\ldots, N$, that is

$$x^{\mu} = \begin{bmatrix} x_1^{\mu} \\ x_2^{\mu} \\ \vdots \\ x_n^{\mu} \end{bmatrix} = \begin{bmatrix} KPI_1^{\mu} \\ KPI_2^{\mu} \\ \vdots \\ KPI_n^{\mu} \end{bmatrix}$$

where μ the observation index and n is the number of parameter types or key performance indices (KPI's) chosen to monitor the environmental condition. In our detection method we use the technique called discrete wavelet transform (DWT) method proposed in [19], which is a mathematical transform that separates the data signal into fine-scale information known as details coefficients, and rough-scale information known as approximate coefficients.

After selecting the data parameters from the data sets, we the produce our experimental databases and calculate the feature weights and averaged it to make a class value. In our technique we use two parameters; one is support threshold θ, and other is correlation threshold δ, in order to decide whether data is normal or abnormal. For instance the temperature is > θ and Humidity is >δ we called this data as malicious data, otherwise the rest of the data is treated as normal. After the threshold calculation we prepare to train the classifier algorithm for evaluation of test data as illustrated in Fig. 6.

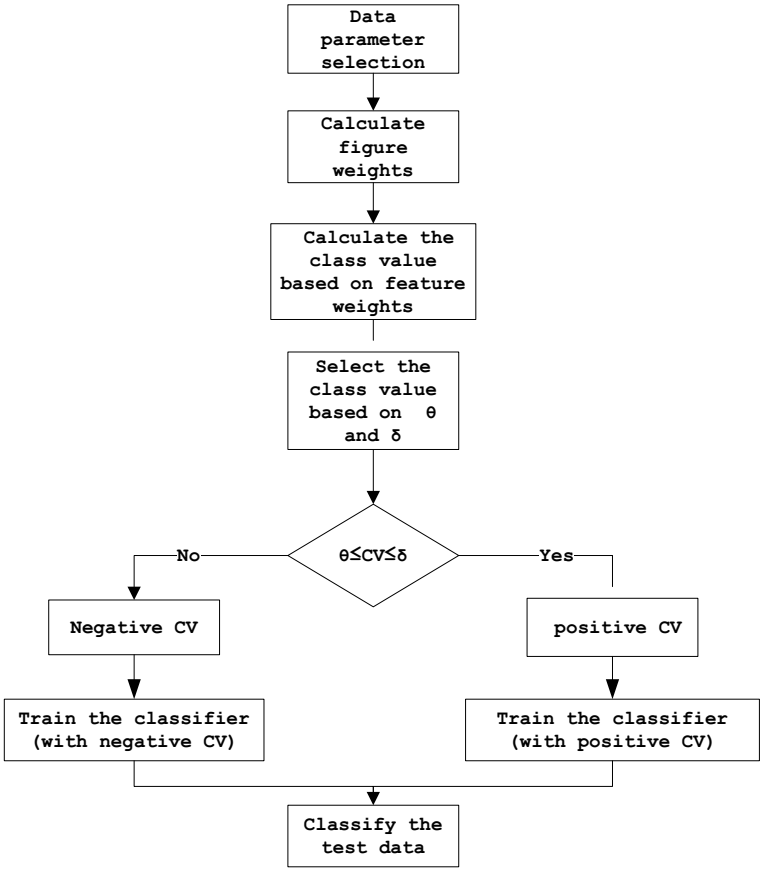


Fig. 6. Detection engine of MDC

5 Experimental Results

The effectiveness of our WSN malicious data detection technique can be measured by the number of FP (false positive) alarms and the TP (true positive) alarms. Since our dataset (IRBL), there are no predefined labels for malicious data, we assessed the data and labelled as malicious data that falls outside the expected value range. In our experiment, we have chosen two parameters, namely temperature and humidity.

Figure 7 shows the average performance of our experiment. In the graph it has been shown that the performance of detection ratio is approximately 95%. It is clear from the graph that the node n11 and n13~n19 shows abnormal behaviour which falls outside of our measured value. Also we can see some of nodes, n46~n51 shows the performance below the measured value. According to our estimation, those node will be identified as malicious node due to their abnormal behaviour of sending data.

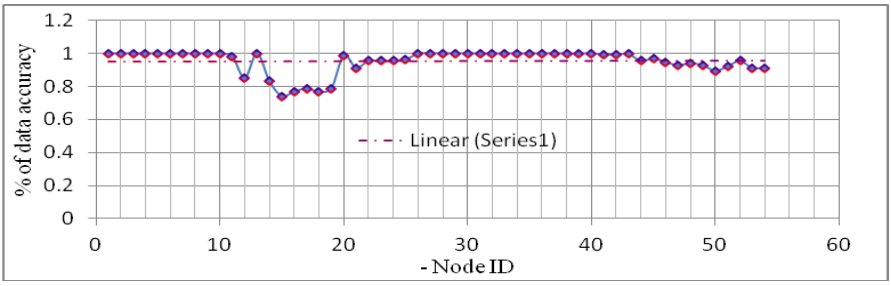


Fig. 7. The average performance of the experiment

The Fig. 8 shows the ROC (Receiver Operating Characteristic) report of our second data set, where we have used five different clusters. The AUC is a popular measure of the accuracy of an experiment. All things being equal, the larger the AUC, the better the experiment is at predicting by the existence of the classification. The possible values of AUC range from 0.5 (no diagnostic ability) to 1.0 (perfect diagnostic ability). The CI option specifies the value of alpha to be used in all CIs. The quantity (1-Alpha) is the confidence coefficient (or confidence level) of all CIs. The P-value represents the hypotheses tests for each of the criterion variables.

Obviously, a useful experiment should have a cut-off value at which the true positive rate is high and the false positive rate is low. In fact, a near-perfect classification would have an ROC curve that is almost vertical from (0, 0) to (0, 1) and then horizontal to (1, 1). The diagonal line serves as a reference line since it is the ROC curve of experiment that is useless in determining the classification. It has been shown from the figure that the abnormality of sensor node falls in second cluster (figure 8.b) and last cluster (figure 8.e). The best cluster is cluster 1 (figure 8.a) and cluster 4 (figure 8.d), which reflects the similar picture presented in figure 7.

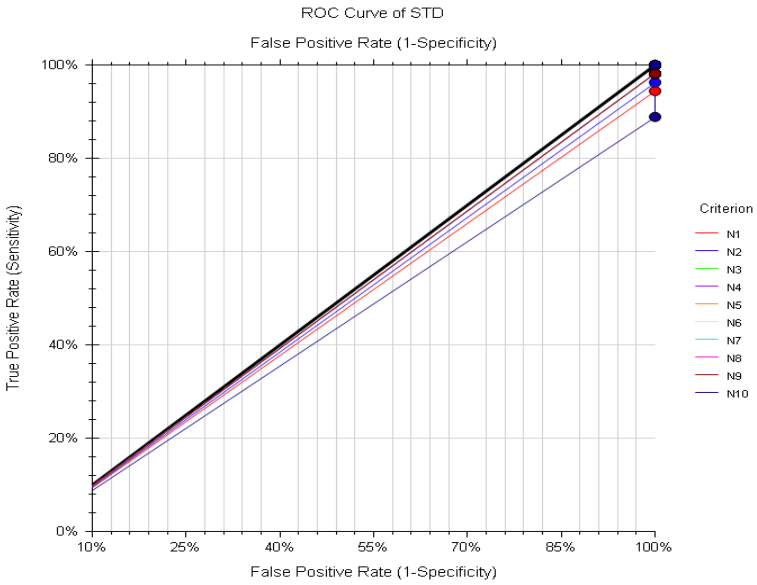


Fig. 8(a). AUC of Chain 1

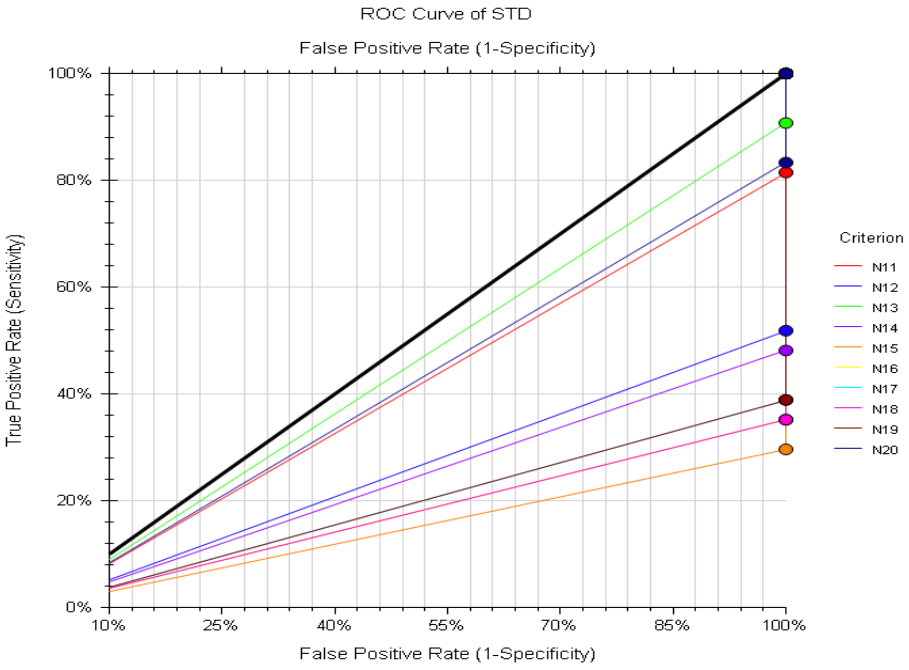


Fig. 8(b). AUC of Chain 2

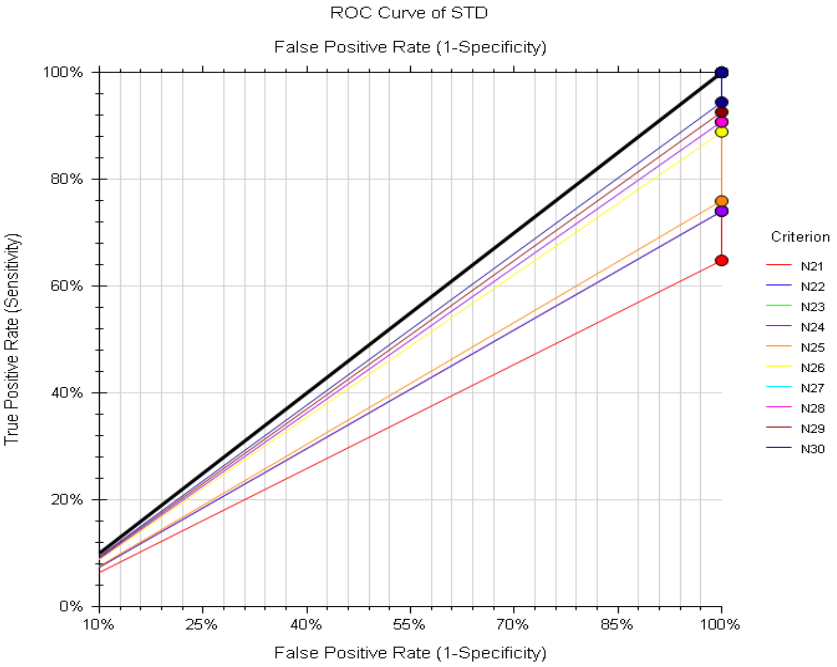


Fig. 8(c). AUC of Chain 3

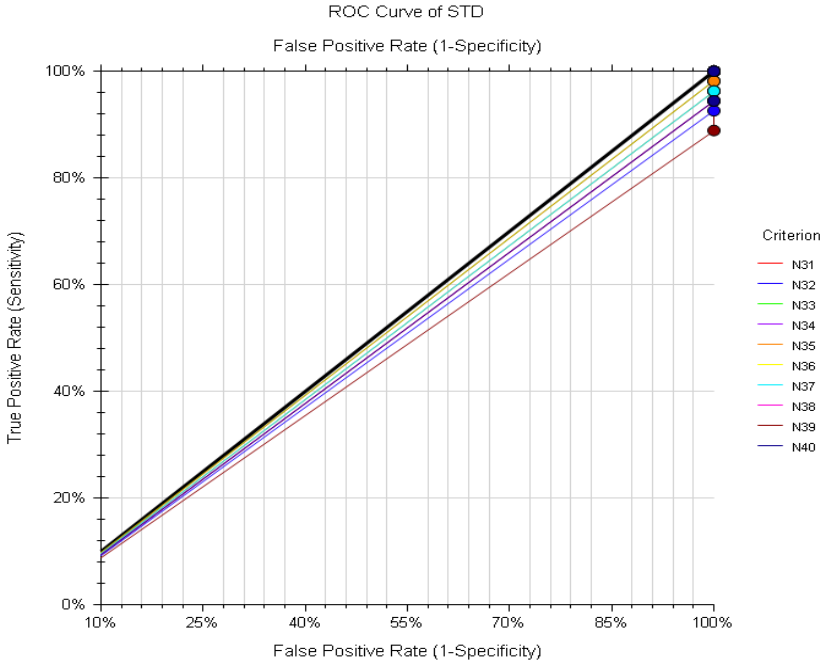


Fig. 8(d). AUC of Chain 4

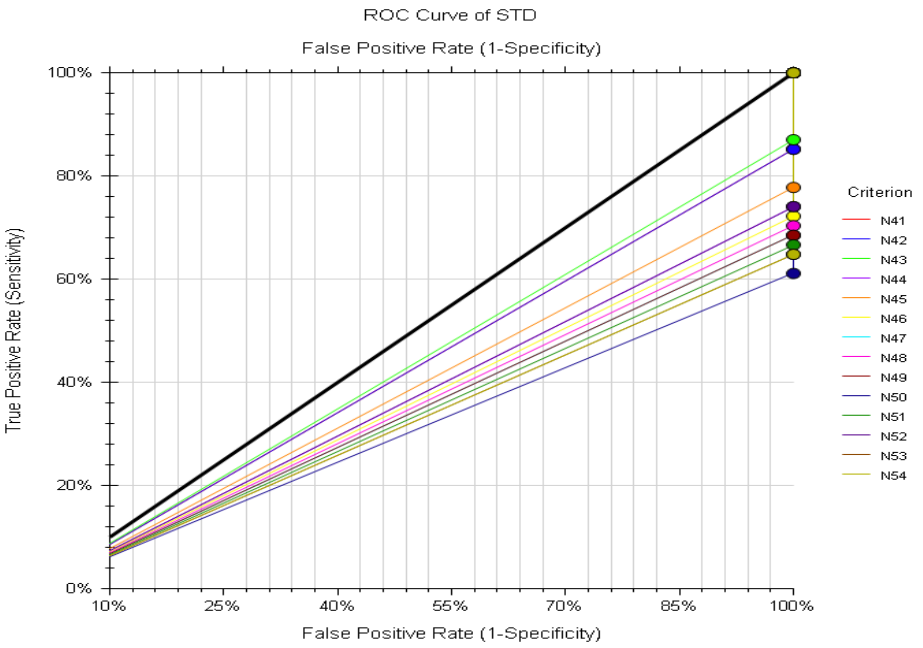


Fig. 8(e). AUC of Chain 5

6 Conclusion and Future Works

This paper proposed a model to identify malicious node from real-world datasets within a non-homogeneous WSN. Our model ensured the data integrity within LN and BS by deploying a detection engine within the MDC. In our simulation, the results show that we can achieve ~70% of detection rates based on our measured value using the real data. In terms of the true alarm rate, the proposed algorithm outperforms. It has been noted that detection ratio has an impact on selecting the threshold value. Our results suggest that the finding optimum threshold can lead to more effective anomaly detection. In particular, our results confirm that the proposed algorithm can maintain acceptable anomaly detection accuracy while using just half of the input data.

In the future, we plan to extend our work to investigate anomaly detection with actual faults obtained from the bioorganic fertilizer plant environment, and study its performance by increasing the DWT level and considering other different types of parameters. Furthermore, we also plan to investigate ways to identify and eliminate erroneous sensor readings at the sensor nodes, which could help further reduce wasted energy from transmitting unwanted erroneous measurements to the base station.

References

1. Karlof, C., Wagner, D.: Secure Routing in Wireless Sensor Networks: Attacks and Countermeasures. Ad Hoc Networks 1(1) (2003)
2. Douceur, J.R.: The Sybil Attack. In: Druschel, P., Kaashoek, M.F., Rowstron, A. (eds.) IPTPS 2002. LNCS, vol. 2429, pp. 251–260. Springer, Heidelberg (2002)

3. Newsome, J., Shi, E., Song, D., Perrig, A.: The Sybil Attack in Sensor Networks: Analysis & Defences. In: The Third International Symposium on Information Processing in Sensor Networks (IPSN 2004), Berkeley, CA (April 2004)
4. Ye, F., Luo, H., Lu, S., Zhang, L.: Statistical En-Route Filtering of Injected False Data in Sensor Networks. In: IEEE INFOCOM 2004, Hong Kong, China (March 2004)
5. Zhu, S., Setia, S., Jajodia, S., Ning, P.: An Interleaved Hop-By-Hop Authentication Scheme for Filtering of Injected False Data in Sensor Networks. In: IEEE Symp. Security Privacy, Oakland, CA (May 2004)
6. Ayman, K., Crussière, M., Hélard, J.-F.: Cross Layer Resource Allocation Scheme under Heterogeneous constraints for Next Generation High Rate WPAN. International Journal of Computer Networks and Communications (IJCNC) 2(3) (2010)
7. Xiao, M., Wang, X., Yang, G.: Cross-Layer Design for the Security of Wireless Sensor Networks. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China (2006)
8. Mamun, Q.: Constraint-Minimizing Logical Topology for Wireless Sensor Networks, PhD Thesis. Faculty of IT, Monash University (2011)
9. Zhao, M., Yang, Y.: Bounded Relay Hop Mobile Data Gathering In Wireless Sensor Networks. IEEE Transactions on Computers 61(2), 265–277 (2012)
10. Chen, Y., Tang, Y., Xu, G., Qian, H., Xu, Y.: A Data Gathering Algorithm Based on Swarm Intelligence and Load Balancing Strategy for Mobile Sink. In: 9th World Congress on Intelligent Control and Automation (WCICA), pp. 1002–1007 (2011)
11. Zhao, M., Yang, Y.: Optimization-Based Distributed Algorithms for Mobile Data Gathering In Wireless Sensor Networks. IEEE Transactions on Mobile Computing 11(10), 1464–1477 (2012)
12. Liang, W., Schweitzer, P., Xu, Z.: Approximation algorithms for capacitated minimum forest problems in wireless sensor networks with a mobile sink. IEEE Transactions on Computers (2012)
13. Zhang, X., Chen, G.: Energy-efficient platform designed for SDMA applications in mobile wireless sensor networks. In: IEEE Wireless Communications and Networking Conference, pp. 2089–2094 (2011)
14. Fei, X., Boukerche, A., Yu, R.: An efficient markov decision process based mobile data gathering protocol for wireless sensor networks. In: IEEE Wireless Communications and Networking Conference (WCNC), pp. 1032–1037 (2011)
15. Zhi, Z., Dayong, L., Shaoqiang, L., Xiaoping, F., Zhihua, Q.: Data gathering strategies in wireless sensor networks using a mobile sink. In: 29th Chinese Control Conference (CCC), pp. 4826–4830 (2010)
16. Ma, M., Yang, Y.: Data Gathering In Wireless Sensor Networks With Mobile Collectors. In: IEEE International Symposium on Parallel and Distributed Processing, pp. 1–9 (2008)
17. Mamun, Q.: A Tessellation Based Localized Chain Construction Scheme for Chain Oriented Sensor Networks. IEEE Sensors Journal 13(7), 2648–2658 (2013)
18. Giannetsos, T., Dimitriou, T., Prasad, N.: Self-Propagating Worms in Wireless Sensor Networks. In: CoNEXT Student Workshop, Rome, Italy (2009)
19. Sharma, K., Ghose, M.: Cross Layer Security Framework for Wireless Sensor Networks. International Journal of Security and Its Applications 5(1), 39–52 (2011)
20. Siripanadorn, S., Hatagam, W., Teaumoroong, N.: Anomaly Detection in Wsns Using Self-Organizing Map and Wavelets. International Journal of Computing 4(3), 74–83 (2010)
21. IBRL-Web, `http://db.lcs.mit.edu/labdata/labdata.html`
22. Mamun, Q.: A tessellation based localized chain construction scheme for chain oriented sensor networks. IEEE Sensors Journal 13(7), 2648–2658 (2013)