Traffic Analysis, a Wise Approach to Detect Compromised Nodes in Wireless Sensor Networks

Mehdi Harizi
Department of Computer Engineering Soosangerd Branch, Islamic Azad University
Soosangerd, IRAN

Mehdi Mohtashamzadeh
Department of Computer Engineering Soosangerd Branch, Islamic Azad University
Soosangerd, IRAN

Aref Saiahi
Department of Computer Engineering Soosangerd Branch, Islamic Azad University
Soosangerd, IRAN

ABSTRACT
Existence of self-similarity and repeated patterns in network traffic makes it possible to predict network parameters. This phenomenon can also be seen in behavior of wireless sensor networks. Self-similarity can be used to propose security solutions in computer and sensor networks. However, sensor networks suffer from several constraints such as limited energy, storage capacity and computing power which make it impossible to take advantage of an overwhelming proportion of tested, evaluated and improved mechanisms which are used in computer networks. Kalman filter is a mathematical tool for estimating control system status with low level of computational overhead. Considering network condition in the past time, Kalman filter can estimate network parameters in future. According to accurate and low cost estimations of Kalman filter, it is compatible with constraints of sensor networks. In this research we have used this concept to present an attack detection mechanism. To this end, Kalman filter make decisions base on traffic volumes produced by different nodes. It can be deduced from results that majority of attacks and abnormal conditions can be marked using the proposed mechanism.

Keywords
Wireless sensor network, Attack detection, Long range dependency of data, Kalman filter.

1. INTRODUCTION
Long-term dependency in network traffic makes it crucial to consider the relationship among data in long time-scales, while this relationship might help to estimate network parameters more precisely. Studies prove the availability of self-similarity in wireless sensor networks. Self-similarity in size of sent packets and arrival time between two consequent packets are evidences of this phenomenon. This unique feature can be exploited to improve the security level of WSNs, so that according to previous knowledge of sensor network status, estimation of parameters of the network in the period ahead is doable, after a while the estimated parameters are compared to actual amounts. If the actual values are very different from the predicted ones, risk of attack in that area of sensor network should be considered further. Precision of the predictions has a significant impact on function of attack detection systems. Kalman filter is a tool for state estimation in control systems. This filter proceeds to prepare future state estimations with regard to the situation in the past and the performed estimates before. Since the Kalman filter can do accurate and low-cost estimations and also considering the limitations of sensor nodes (limited power and processing resources), this filter can be used for attack detection in sensor nodes.

This research has tried to introduce a mechanism to estimate the volume of generated traffic to detect occurrence of attacks. The presented architecture has been implemented and evaluated in omnet++ simulator. Results depict that in case of exploiting presented architecture an overwhelming proportion of attacks and abnormal variations of network parameters are traceable. However, in some cases Kalman filter estimations differ from actual status of sensor nodes. This could be due to inaccurate predictions of filter or changes in network status (compared to previous traffic patterns).

2. Self similarity in wireless sensor networks
In [7] a network which is included six sensor nodes and a base station was considered. Connection of sensor nodes to base station directly and once again the connection was considered to be daisy chain. Network traffic in both topologies has been studied for three times and each time 24 hours. The authors measured two main parameters, size of sent packets and Inter-arrival time of packets, and came to this conclusion the second parameter has self similar behavior.

Authors of [3] have proved that received data from sensor nodes are self similar and predictable. The proposed mechanism in their research has two parts, “signal predictor” and “Event occurrence predictor”. Their proposed method does predictions by the mean of a Fuzzy controller.

In [1], assuming the existence of self similarity in network traffic, author has proposed a mechanism to estimate the rate of self similarity by which Queue management process and quality of service have been improved. Achieving shorter queue length and response times show the accuracy of assumption of availability of self similarity in network traffic.

Authors of [2] have divided data transmission of sensor networks to three categories:

1- Query Driven: in this case a supervisor node informs all other nodes about data type which it is interested to receive. From now on sensors send packets related to the events which are in interest domain of a specific supervisor node.

2- Event Driven: In this case each sensor node sends events' packets to all supervisors immediately after receipt. In this type, sensor nodes commence data transfer process. This kind of transmission may lead to burst traffic and repeated data.

3- Continuous Driven: In continuous transmission, sensors and supervisors all can commence data transfer, but usually transmission is done in rhythmic and predefined periods of time from sensors to supervisors. If data have high priority, parameters such as bandwidth and delay are of importance. In contrast in transmission of data with low level of priority, reliability parameter is important.

Authors have concluded that Event driven and Query driven types have self similarity and cannot be modeled by Poison model.

3. Abnormal function detection algorithm (Attack detection algorithm)

As mentioned before, whole traffic in sensor networks may have repeated patterns in long time-scales. This self similarity is as a result of repeated traffic patterns generated by each individual node. Simulation results show the existence of self similarity for average sending rate parameter for some long time-scales (especially time-scales larger than 2.5 seconds). So amount of packets which are going to be transferred in the time period ahead can be estimated. Comparing this estimation with the actual sending rate at the end of the time period, algorithm can find whether a node functions normally or not. If the difference between estimated and actual values is great, the node is suspected to be compromised. However, ensuring that a node has been really compromised needs deeper considerations (considering function of nodes in more than one period of time and asking other nodes which are deployed in that area), while the increase in sending rate may be because of occurrence of a specific event (detection of fire, light and etc).

Since the Kalman filter can provide accurate and low-cost estimations [6], and also considering the limitations of sensor nodes (limited power and processing resources), this filter can be used to detect attacks in wireless sensor networks.

The proposed mechanism functions as follows: several nodes are randomly chosen as supervisor nodes by the base station for a defined amount of time. Each supervisor is responsible for estimating the amount of traffic (which is going to be generated) by every deployed node in its corresponding region for the time period ahead. Then during this time period, supervisor records the amount of traffic emitted from all other nodes in its region and average size of sent packets. If the difference between estimated and actual values is more than the pre-defined threshold, the probability of occurrence of an attack should be considered. In this research, comprehensive consideration of attack occurrence includes three main actions:

Assess the amount of traffic generated by nodes adjacent to suspected node.

Do estimations and comparing with actual values in the next time-scale.

Comparing average sent packets in the next time-scale with pre-stored information.

If the increase in generated traffic is as a result of detection of an event in the area covered by the sensor node, the neighbor nodes must also have experienced such increase. Therefore, first consideration can properly differentiate normal and attack occurrence mode, in most case. If adjacent nodes do not show noticeable increase in traffic volume, the second action may examine the possibility of occurrence of an attack more precisely. In this phase, the third action may help to increase the certainty of the final decision. The third action is based on the concept that attacker tries to maximize sending rate and consuming power of the compromised and receiver nodes.

4. Using Entropy to determine long time-scale [4]

Entropy is a tool for the study of hidden information in a probability distribution. For the probability distribution function Pi entropy is calculated as follows:

$$S(P_i) = \sum_i P_i \log \frac{1}{P_i}$$

In this research the probability distribution function is $P(R_2 = \acute{r} | R_1 = r)$ in which 'r' and 'ŕ' are equal to mapped values of generated traffic to numbers 1,2,3, or 4. In other words,

if "generated traffic" $\in (0 , 2.5]$ "packet/sec" then r=1

if "generated traffic" $\in (2.5 , 5]$ "packet/sec" then r=2

if "generated traffic" $\in (5 , 7.5]$ "packet/sec" then r=3

if "generated traffic" $\in (7.5 , 10]$ "packet/sec" then r=4

Therefore, entropy can be computed for this bi-conditional probability distribution function as follows:

$$S = \sum_{\acute{r}} P(R_2 = r' \mid R_1 = r) \log \frac{1}{P(R_2 = r' \mid R_1 = r)}$$

If the probability distribution is uniformly continuous, entropy is maximum and if it is concentrated at a certain point, entropy would be minimum. For $P(R_2 = r' | R_1 = r)$ since transmitted traffic is self similar, the probability should be concentrated to 'r'.

Fig. 1 shows the computed values of entropy for various long time-scales. As it can be seen in this figure, entropy has a low value for 2.5 seconds long time-scales. Although for some larger time-scales entropy is less, choosing 2.5 second long time-scale seems to be appropriate since selecting very large time-scales would decrease the precision of estimations and increase the delays caused by false estimations.

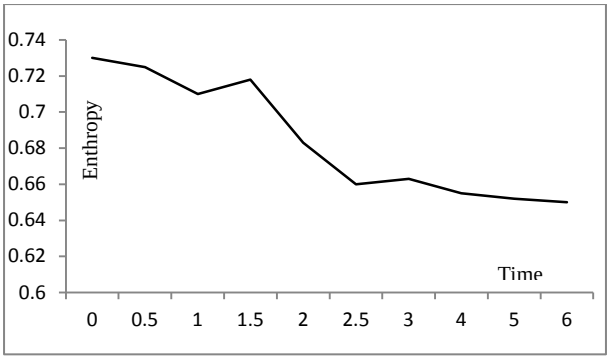

Fig 1.computed values of entropy for various long time-scales

Table 1.presents the probability $P(R_2 = \acute{r} | R_1 = r)$ for various values of r and ŕ after 300 seconds.

Table 1.P $(R_2 = \acute{r} | R_1 = r)$for various values of R_1 and R_2

R_1 Level(pkt/sec) \ R_2	1	2	3	4
1 { (0,2.5] }	0.58	0.31	0.08	0.03
2 { (2.5,5] }	0.26	0.47	0.21	0.06
3 { (5,7.5] }	0	0.15	0.64	0.21
4 { (7.5,10]}	0	0.06	0.38	0.56

As it can be deduced from Table 1, R_2 (generated traffic during the time-scale ahead) is close to R_1(generated traffic in the last time-scale) in most cases. As an example, for $R_1 =$ 3(third row), R_2 is equal to 3 in 64% of situations. Furthermore, R_2 has been equal to 4 and to in 21% and 15% of situations, respectively and it never has been 1. It is evident that R_2 is concentrated to 3 when R_1 is equal to 3. This situation is perceptible in other rows in the table.

5. Simulation results

Simulations have been done by the mean of omnet++ simulator with the following parameters:

Radio propagation	Two Ray ground
Routing protocol	AODV
MAC protocol	802.15.4
Interface queue	Drop tail
Packet size	512

Fig. 2 (a and b) show two different parts of sensor network after 450 seconds. In both cases, supervisors are marked by ⊙.

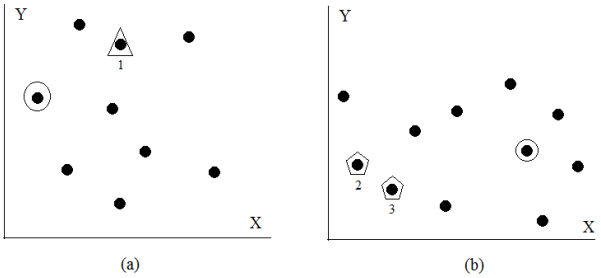

Fig 2.Two different parts of simulated network after 450 seconds

In fig. 2 sensors which are marked with ⊙were detected by supervisors as compromised nodes (in the initial considerations), for which more considerations (three introduced actions mentioned earlier) have rejected this assumption. On the other hand the node which is marked with △ was assumed by corresponding supervisor as compromised node for which more considerations have proven it.

Fig. 3, Fig. 4, and Fig. 5 presents the computed values for generated traffic estimations by the nodes number 1, 2, and 3 and the actual values of generated traffic during three time-scales before an three time sales after 450 seconds. In all cases, nodes are initially were assumed to be compromised since the difference between actual and estimated vales of generated traffic were more than one level.

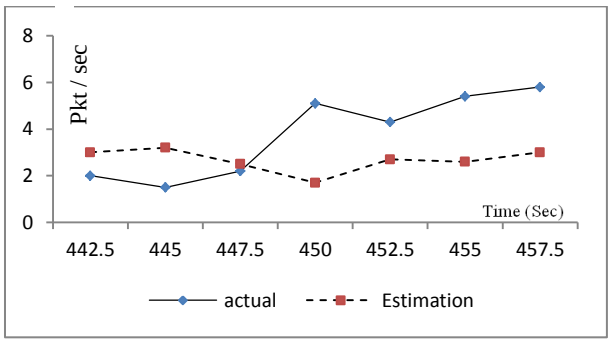

Fig 3.Estimated value of generated traffic Vs. actual value for node number 1.

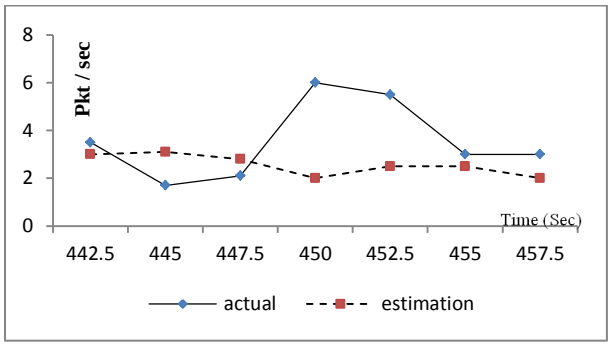

Fig 4. Estimated value of generated traffic vs. actual value for node number 2.

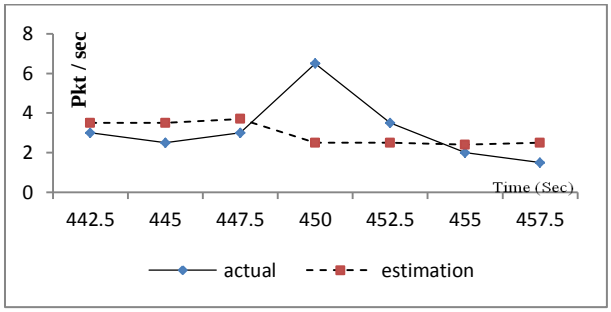

Fig 5. Estimated value of generated traffic vs. actual value for node number 3.

As it can be seen in Fig. 3 the difference between estimated value of Kalman filter (level 1) and the actual value (level 4) is more than one level at 450 seconds after simulation starts, the supervisor assumes that the node number 1 might be compromised. During simulations, polling from adjacent nodes (action 1) depicted that no event has been detected by others in that specific time-scale. furthermore, the stability of amount of generated traffic by this node in the next time-scale and the difference with the estimated value in this step (action 2) and also the difference between the volume of sent packets in comparison with previously stored information, confirms occurrence of an attack to node 1.

It can be extracted from the results that node 2 has experienced a growth in traffic volume at 450 seconds compared to its previous time-scale (447.5 seconds). This has led to have a gap between the actual amount of generated traffic (6 pkt/sec) and the value predicted by Kalman filter (2 pkt/sec) at the time of 450 seconds. Although this gap was still evident in the next time-scale, feedbacks from other nodes deployed in that area indicated detection of occurrence of an event and as a result experiencing increase of sending rate for most sensors.

As it can be seen in Fig. 5 the node 3 (which is adjacent to node 2) has also experienced such increment at the time of 450 seconds. However, while during the next time-scale (452.5 seconds) the estimated and actual values of generated traffic has been in the same level (action 2) Likelihood of occurrence of attack is ruled out.

Simulation results have been investigated for all nodes. Studies show that in most cases the proposed algorithm performs accurate in detection of attacks. However, there might be some contradictions between filter estimations and actual status of sensor nodes which can be due to network status changes (compared to previous traffic patterns).

It is worth noting, if events in the network conforms a repeated and self-similar pattern, estimations might be more precise.

6. Conclusion

Self-similarity and existence of repeated patterns are inseparable characteristics of networks. This phenomenon can be seen in wireless sensor networks. Although there are lots of common aspects between sensor and computer networks, proposed mechanisms for computer networks in various fields (traffic engineering, quality of service, security and etc.) can hardly be used in sensor networks due to some constraints, namely resources of power and processing and memory. These limitations should be considered carefully in case of implementation of security solutions since this category of mechanisms are of the high priority and must be performed in all sensor nodes. In this research the focus has been on using tools by which network status could be estimated and predicted. However, most of prediction tools such as data mining and neural networks exert complicated computations and need large amount of memory resources. In contrast, Kalman filter, which is used to estimate status of control systems based on its understandings from systems in the past and estimations which has been done before by it, needs small amount of resources and is compatible with the constraints mentioned above. In this work the attempt was to present a mechanism which takes advantage of Kalman filter to estimate the volume of generated traffic by sensor nodes and use this information to detect occurrence of attacks.

The proposed mechanism was evaluated using omnet++ simulator. Results depicted that majority of attacks and abnormal status changes in sensor networks can be detected by this mechanism. However, there were reasonable level of uncertainty and false predictions which are an inseparable part of every control system.

7. ACKNOWLEDGMENTS

The authors gratefully acknowledge the Research Office of Irrigation and Drainage Networks of Khuzestan Water and Power Authority (KWPA) for their financial support.

8. REFERENCES

[1] L. Heng and W. Yan. An adaptive proportional integral active queue management algorithm based on self-similar traffic rate estimation in WSN, Journal of Korean Society for Internet Information, Vol. 5, No. 11, 2011.

[2] Y. Jiao Wang and H. Yun Lin. A Kind of Improved RED Algorithm of WSN Oriented to Self-Similar Traffic, Advanced Materials Research, Vol. 214, 2011.

[3] Q. Liang. Energy Efficient Wireless Sensor Networks Using Fuzzy Logic, Bi-annual (12/2004-05/2005) Performance/Technical Report for ONR YIP Award, 2005.

[4] K. Park and W. Willinger, SELF-SIMILAR NETWORK TRAFFIC AND PERFORMANCE EVALUATION, John Wiley, 2000.

[5] K. Park, and T. Tuan, "Performance evaluation of multiple time-scale TCP under self similar traffic condition," ACM transaction on modeling and computer simulation, Vol. 10, No. 2, pp. 152-177, April 2000.

[6] B. Sun, L et al. Integration of Secure In-Network Aggregation and System Monitoring for Wireless Sensor Networks. IEEE ICC '07, Glasgow, U.K., June 2007.

[7] W. Wei Beng. ANALYSIS AND CLASSIFICATION OF TRAFFIC IN WIRELESS SENSOR NETWORKS, PhD thesis, Naval postgraduate school, Monterey, California,2007.