A Method for Countering Snooping-Based Side Channel Attacks in Smart Home Applications

Jingsha He, Qi Xiao(✉), and Muhammad Salman Pathan

School of Software Engineering, Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing 100124, China
jhe@bjut.edu.cn, xqnssa@emails.bjut.edu.cn,
muhammad.salman@nu.edu.pk

Abstract. In recent years, with the rapid development of the Internet of Things (IoT), the information technology has been widely used in smart home applications. On the other hand, smart home technology closely related to people's privacy, which is not much considered by smart home vendors, making the privacy protection of smart home a hot research topic. Traditional encryption methods can ensure the security of the transmission process, but it can hardly resist the side channel attacks. Adversaries can analyze the radio frequency signals of wireless sensors and timestamp series to acquire the Activity of Daily Living (ADL). The most simple and efficient way to counter side channel attacks is to add noise into the transmitted data sequence. In this paper, we propose an improved method based on Logistic Regression (LR), which can be adapted to network status to protect the privacy of residents in smart home environments. Compared with other similar approaches, our method has the advantage of low energy consumption, low latency, strong adaptability and good effect of privacy protection.

Keywords: Smart home · Side channel attack · Privacy · Logistic Regression

1 Introduction

Smart home is one of the important branches of the Internet of Things (IoT) which relies on wireless sensors to sense and collect activity and status information. These sensors can sense particular phenomena, convert the sensed information into data, process the data and then transmit the data onto a sink node for further analysis [1]. For example, the measurement of temperature, humidity, luminosity, noise levels, presence, etc., can provide useful data to interpret a physical activity in space and time in order to determine the activity of a person and thus can contribute in detecting unusual situations and emergency cases [2]. The sensed data contains much private information of the resident, however, while the societal concerns of smart home technology evolution in relation to the privacy and security of the citizen appear to be still at an embryonic stage [3]. The acquired ADLs can help to improve the quality of life, but it can also be exposed to the attacker, therefore, the issue of privacy protection in smart home environments has become one of the most challenging issues.

Q. Chen et al. (Eds.): ChinaCom 2016, Part I, LNICST 209, pp. 200–207, 2018.
DOI: 10.1007/978-3-319-66625-9_20

Compared to other wireless sensor networks, the type and number of sensors in the smart home are similar to general homes. And most sensors are operated in the event-triggered mode, where sensor data is transmitted only when a relevant event is detected. We can image that when an event occurs, the transmission will be triggered immediately. While data encryption algorithms can only ensure the security of the sensor data during the transmission, however the radio frequency of the transmission can be revealed to the adversary who has the ability to listen to the global transmission state. The adversary can use a side channel attack method to analyze the transmission sequence and can acquire the ADLs of the resident. For example, Fingerprint And Timing-based Snooping (FATS) attacks only need the timestamp and the fingerprint of each radio message, where a fingerprint is a set of features of an RF waveform that are unique to a particular transmitter [4]. The most simple and effective method to resist side channel attacks is to add fake message onto the transmission sequence to make the adversary unable to distinguish between fake and real messages. However, due to the limitations of the communication bandwidth, battery energy and computing power of wireless sensor nodes, the amount of fake data should be added as low as possible.

There has been an extensive study of the approaches to protect the privacy of residents in a smart home environment. These solutions are mainly based on a fixed frequency or probability models, thus having the major drawbacks like delaying in reporting the real events until the next scheduled transmission. For smart home scenarios, such delay of reporting real sensed data can cause the degradation of the quality of service (QoS) in many applications [5]. Some of the applications like intelligent sensing, the delay can generally not be tolerated, where the states of the sensors must be received in a real timely fashion for making the corresponding responses. As the delay problem is concerned, Park et al. proposed an improved method based on behavioral semantics. But the method heavily depends on the accuracy of prediction, if the prediction of the next activity provides an inaccurate answer, the added fake messages will not be enough to affect the statistical analysis. In this paper, we propose an improved method to resist the side channel attacks based on logistic regression that can be adapted to the network state. That is, when the traffic is heavy and no ADL happens, the frequency of adding fake data should be automatically reduced. When the real event occurs, right time will be chosen to add noise to protect the real events.

The rest of this paper is organized as follows. In Sect. 2, we review some existing solutions. In Sect. 3, we describe our method in detail. In Sect. 4, we will compare our method to some other solutions. In Sect. 5, we have given the conclusion of this approach.

2 Related Work

The ConstRate (Sending packets at a fixed frequency) model, all the sensor nodes send the packets according to the same transmission intervals. Thus, the real events must send the packets until the next transmission. So the method can achieve the remarkable work to resist the static analysis attack. Obviously, the ConstRate model has a congenital deficiency: the delay depends on the transmission interval which is half of the interval. Also it is difficult to determine an appropriate transmission interval in the

ConstRate Model. The delay will vary with the time interval. When the time interval increases, the delay time will also increase. And the amount of fake messages and additional energy consumption will be increased significantly.

Shao et al. proposed the FitProbRate model that aims to improve the deficiencies of ConstRate model. The core idea is to make the intervals follow the specific probability distribution. When a real event occurred, the algorithm will start looking for a minimum interval which obey the distribution of index distribution to send. When the real event is sparse, the FitProbRate model will get a good performance, and compared with the ConstRate model the delay will be reduced. On the contrary, the real event triggered frequently, the delay will bigger than other models.

Park et al. proposed a model which based on adding several fake packets to the transmission sequence [6]. The model adds fake packets lie on the events that will happen in the future. The first step is to forecast the activity through the status of the sensor nodes and then the fake messages will be generated according to the prediction. Even an attacker has the ability to listen to the transmission of all the sensors, it could only predict the wrong ADLs. However, the shortcoming of the model is that the effect depends on the answer of the prediction. If the prediction model gives a wrong forecast, the fake packets will pall on the protection of the ADLs. Obviously, the stability of the model is lower than these two models mentioned above.

The purpose of adding fake packets to the transmission sequence is to make the attacker can't pick out the fake packets from all the RF radios. For the top two models above, the interval of all the packets is obey to the same distribution. It is assumed that the attacker have the ability to listen to the RF radio of the whole wireless sensor network. And make the model effectively, the transmission sequence must have the significant confidence to make sure that the adversary couldn't determine the real radio is contained in which intervals [7]. That is, if the transmission sequence which contains the fake packets in sending real messages has the sufficient randomness, the adversary cannot recognize the fake data from the real messages, and the ADLs of the residents will be protected. In our method, we have made enough randomness between fake and real data to ensure the attacker couldn't recognize the fake messages. As for the load of traffic is concerned, the more closer to the sink node, the larger data is needed to forward. If all the sensors send the packets follow the same distribution, the sensor node near the sink node will be too heavy to forward packets. Therefore, it is necessary to make the sensor node sends fake data packets adaptively according to the network status.

Considering the particularity of the wireless sensor network and the sensitivity of the smart home, it is important to think over the privacy, energy consumption and latency of the WSN while designing the noisy based privacy protection models. For the effect of privacy protection, the noise data should not be recognized, but also makes the correct recognition rate of the behavior low enough. In other words, either the identification of the behavior should be wrong, or can't recognize the true behavior. As for energy consumption, it cannot be a good privacy protection model, if the implementation of the model greatly reduces the lifetime of the WSN. We should consider the average of traffic load to prolong the lifetime of the sensor network. Latency is the main indicator of the QoS. If the latency is too long, it can lose the meaning of intelligence. As for the delay, a good model should make the delay as small as possible. Especially

in the automatic adjudication environment, the sink node use the long delayed messages is meaningless.

3 Proposed Method

In this section, we have introduced our method. The logistic regression will be used to judge whether the fake messages should be sent or not, and named as hypothesis function. At first, the sensor node will acquire the current state, and normalize the data by a simple process. Then, input the processed data to the hypothesis function to decide whether to send the fake message. For each sensor node, the parameters of the hypothesis function will be different. And the different parameter will be trained by the sink node through supervised learning.

At the very beginning of our method. All of the sensor nodes in the sensor network will send fake messages with a fixed time-window. The detailed procedure is described in Fig. 1. The sensor node needs three variables to plug into the hypothesis function, to determine whether the fake data should be sent or not. These three variables are traffic state, time and send density, represented by x_1, x_2, x_3. We collected the traffic state and send density in the period of the block time. We present the real-time transmission times as TT, which consists of the send times of themselves and the forwarding times from other sensor nodes. And we use TT to divide by TrMax to represent the current traffic situation, where the TrMax is the max transmission times of all the sensor nodes in the block time. In order to unify the time, we map the current timestamp to the region of [0, 24], and presents as x_2. As for the send density, we use the Send Times (ST) to be divided by BlockMax to represent the send density, where the BlockMax is the block interval divided by time-window. When these three parameters are collected, the ANS could be calculated by the hypothesis function. If the ANS is greater than 0.5, fake

Algorithm 1: Sensor node checks for sending fake messages
Repeat
$x_1 = TT/TrMax$;
$x_2 = (nowtime - date(now))/3600$;
x_3 = ST/BlockMax;
$ANS = 1/(1 + exp(-\theta^T x))$
if $ANS \geq 0.5$ **then**
    wait (rand (0, window));
    send ({sensor_id : sensor_id, traffic_load: x_1,time: x_2,is_fake:1});
    fail_fake = 0;
    window = initial window;
  else
    fail_fake = fail_fake+1; //record does not send fake data times
    if $fail\_fake \geq max\_times$ **then** window =2*window;
    //the time window becomes double
end if
until next window

Fig. 1. Sensor node checks for sending fake messages

massage should be sent after a random time interval. Otherwise the algorithm will judge whether to send the messages at long intervals, if so, the time-window would be increased. Once the real data is sent, the time-window will be recovered again.

Figure 2 Shows the detail of the learning algorithm. Each sensor node has its own different parameters, the learning algorithm will deal with all the data and calculates θ for each sensor node respectively. As for each dataset from different sensor nodes are concerned. Before the learning algorithm, the received data should be labeled. The main purpose of labels is to mark the fake data, labeled with 1 indicates that fake data should be sent, labeled with 0 indicates that the fake data should not be sent. As for the learning algorithm, Firstly, we calculate the cost by using the square difference method, just as the J(θ) in the Algorithm 2, the m in J(θ) is the number of training samples. And then the gradient descent algorithm is used to find the θ that makes the J(θ) obtain the global minimum [9]. When the learning algorithm is completed, the θ should be handed out to the sensor node and should be plugged into the hypothesis function.

Algorithm 2: Sink node training the sensor data
Input: the set of all sensor data X
Output: θ
Foreach X as X_i
Label the x_i *by* traffic status
End Foreach
Foreach X as X_i
$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(X^{(i)}\right) - y^i\right)^2$
Repeat
$\theta_j^{(i)} := \theta_j^{(i)} - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$
Until Convergent
End Foreach
Hand out all parameters

Fig. 2. Sink node training sensor data

4 Evaluation

In this section, we will first introduce the method and settings of the experiment. And then will do the experiment to compare the effect of privacy protection and delay with the ConstRate Model and FitProbRate Model. We have used the public dataset related to accurate activity recognition in a Home Setting [8] for our experiment.

In order to evaluate the effects of the privacy, we studied the side channel attack method in a smart home. The common step of the side channel attack is to cluster the sensor data to the reasoning for the number of rooms in the home. The purpose of the FitProbRate model and ConstRate model are to disturb the attacker to cluster the sensor to a wrong classification. In our experiment we use the cluster accuracy to evaluate the protection model [5]. If the clustering accuracy is approximate to 1, the clustering results will equal to the number of rooms and the sensor distribution is the same as the

clustering results. In contrast, if the clustering results are completely inconsistent with the actual room distribution, the ADL will be perfectly protected. Consequently, the lower the cluster accuracy will be, the better effect of privacy protection will be achieved. In addition, the Ratio is calculated by the number of fake messages divided by the number of real messages.

As shown in Fig. 3, with the incensement of the Ratio, the clustering accuracy of these three model decreased gradually. The clustering accuracy of the ConstRate model maintained at a low level. Our method and the FitProbRate model are affected by the Ratio, because when the Ratio is increased, the clustering accuracy is declined, And when the ratio is in the range of (5, 15), the clustering accuracy between our method and FitProbRate model has a wide margin, and the FitProbRate model has a good performance than our method. When the amount of fake messages is low, the FitProbRate model has been well-distributed than our method. So our method is lower in accuracy than the FitProbRate model in terms of clustering accuracy in the range of (5, 15). When the Ratio is greater than 15, the gap between our method and FitProbRate model is very small. Also, we have ensured that when the clustering accuracy is lower than 0.4, the side channel attack will be hard to analyze the ADL.

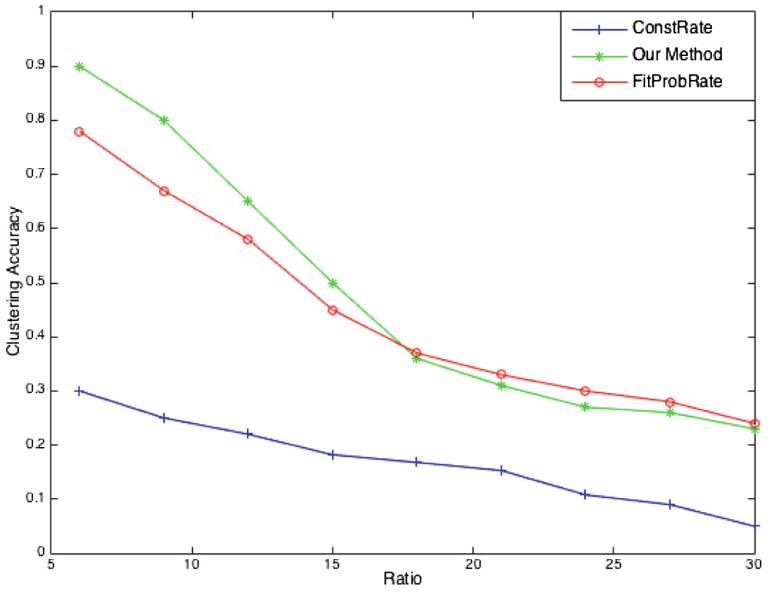

Fig. 3. The relation between ratio and clustering accuracy

As shown in Fig. 4 with the incensement of the Ratio, the latency of ConstRate model and FitProbRate model decreased gradually, and the latency of the ConstRate model is the longest of the three models or method, which is the half of the transmission interval. Compared with the ConstRate model, the FitProbRate model has significantly declined the latency. Figure 4 displays the average of the latency of these

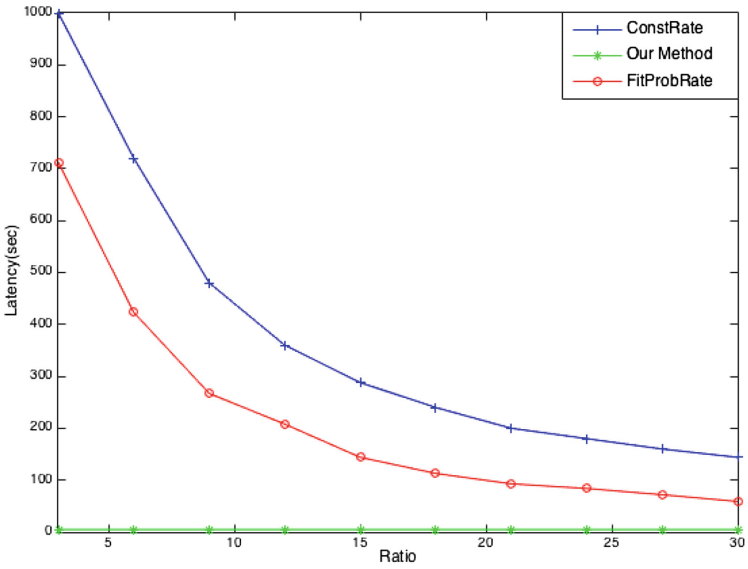

Fig. 4. The relation between ratio and latency

three model, the latency of the FitProbRate model related to the density of the transmission, when the transmission frequency of the real event is raised up, the latency will be increased. On the contrary, when the transmission is sparse, the latency time will be declined. Consequently, the latency of the FitProbRate model is affected by the frequency of the real events.

5 Conclusion and Future Work

In this paper, we have proposed a new method to resist the side channel attack. Compared with other models, our method has the advantage of adaptive network status, low latency and low power consumption. In the context of the smart home environment, when the user goes to work and when comes back to home from work or when go to sleep etc., these are likely to be the scope of privacy protection. The attacker can infer these living habits through analyzing the density of the transmission sequence. In the future, we plan to make our method fit the daily routines to the popular routine of almost people. Even if the attacker gains the daily routines, also cannot distinguish the routines of the particular prey. In a word, the privacy plays an essential role in the smart home, we should pay more attention to the privacy protection in smart home.

Acknowledgement. The work in this paper has been supported by Beijing Natural Science Foundation (4142008), National Nature Science Foundation of China (61272500) and National High-tech R&D Program (863 Program) (2015AA017204).

References

1. Theoharidou, M., Tsalis, N., Gritzalis, D.: Smart Home Solutions: Privacy Issues. Handbook of Smart Homes, Health Care and Well-Being, pp. 1–14. Springer, Cham (2014)
2. Alami, A., Benhlima, L., Bah, S.: An overview of privacy preserving techniques in smart home wireless sensor networks. In: 10th International Conference on Intelligent Systems: Theories and Applications, pp. 1–4. IEEE Press, Rabat (2015)
3. Sanchez, I., Satta, R., Fovino, I.N., Baldini, G.: Privacy leakages in smart home wireless technologies. In: International Carnahan Conference on Security Technology, pp. 1–6. IEEE Press, Rome (2014)
4. Srinivasan, V., Stankovic, J., Whitehouse, K.: Protecting your daily in-home activity information from a wireless snooping attack. In: International Conference on Ubiquitous Computing, pp. 202–211. ACM Press, Seoul (2008)
5. Park, H., Park, T., Sang, H.S.: A comparative study of privacy protection methods for smart home environments. Int. J. Smart Home **7**, 1–12 (2013)
6. Park, H., Basaran, C., Park, T., Son, S.H.: Energy-efficient privacy protection for smart home environments using behavioral semantics. J. Sens. **14**, 16235–16257 (2014)
7. Yang, Y., Shao, M., Zhu, S., Cao, G.: Towards statistically strong source anonymity for sensor networks. In: 27th Conference on Computer Communications. pp. 51–55. IEEE INFOCOM, Phoenix (2008)
8. Van Kasteren, T., Noulas, A., Englebienne, G.: Accurate activity recognition in a home setting. In: International Conference on Ubiquitous Computing, pp. 1–9. ACM Press, Seoul (2008)
9. Song, Y., Cai, Q., Nie, F., Zhang, C.: Semi-supervised additive logistic regression: a gradient descent solution. J. Tsinghua Sci. Technol. **12**, 638–646 (2007)