Trace malicious source to guarantee cyber security for mass monitor critical infrastructure

Trace malicious source to guarantee cyber security for mass monitor critical infrastructure

Xiao Liu[a], Mianxiong Dong[b], Kaoru Ota[b], Laurence T Yang[c], Anfeng Liu[a,*]

[a]School of Information Science and Engineering , Central South University, ChangSha, China

[b]Muroran Institute of Technology, Japan

[c] Francis Xavier University Antigonish, Canada

Abstract: The proposed traceback scheme does not take into account the trust of node which leads to the low effectiveness. A trust-aware probability marking (TAPM) traceback scheme is proposed to locate malicious source quickly. In TAPM scheme, the node is marked with difference marking probability according to its trust which is deduced by trust evaluation. The high marking probability for low trust node can locate malicious source quickly, and the low marking probability for high trust node can reduce the number of marking to improve the network lifetime, so the security and the network lifetime can be improved in TAPM scheme.

Keywords: Cyber security, cyber forensics, traceback, trust, marking probability, network lifetime

1 Introduction

Our society, economy and critical infrastructures have largely depended on information and communications technologies (ICT) [1-6]. Cyber-attacks [2-6] are becoming more attractive and can lead to large-scale (or global) systemic failures, resulting in loss of human life and social unrest with our dependence on information technology increases. Wireless sensor networks (WSNs), as one of pivotal component of Cyber-Physical Systems (CPS) [7-8], plays an irreplaceable role in roads, railways, industrial oil pipeline, and a wide range of environmental monitoring [3, 4, 7-13].

Sensors are often highly cost-sensitive. They require smaller processors and memories which are resource-constrained in energy due to the nature of small size [1, 7, 11-13]. They are often unattended and prone to different kinds of attacks because of their operating nature [9-12, 14]. For example, DDos attack, Ref. [15] proposed a scheme to resist this attack. In order to ensure data security, Ref. [16] present a new SCA-WSN scheme that not only achieves user anonymity but also works with the computation loads for sensors effectively. As one of cyber forensics technologies, traceback is a promising solution to counter the cyber-attacks by determining the probable source of malicious node. Packets marking approach is an effective traceback approach for malicious attack. In order to locate malicious source, the victims consult upstream nodes to reconstruct attack paths by broadcasting the information of the malicious packet(s) in the traceback request [10, 11, 12, 13].

In Packets marking scheme [10, 12], each node adds its ID information to data packets in the routing process, the marking information will be longer with the routing of data packets, which can damage the network lifetime. To weaken the effect of Packets marking scheme on the network lifetime, probability marking scheme (PM) is proposed in Ref. [10]. Each passed packet are marked by the nodes with a certain probability, which can reduce the amount of mark information and improve the network lifetime. In traceback scheme, the key method is to marking generated data packets by suspicious nodes, thus it can provide many useful information for determining malicious source, and the packets generated by the "good" nodes are not marked in the network, which can't damage the network lifetime. In previous schemes, the marking probability have nothing to do with the nodes' credibly. Thus the performance for

* Corresponding author.
E-mail address: afengliu@mail.csu.edu.cn (Anfeng Liu).

system is not good enough.

Based on the above analysis, this paper proposes a new trust-aware probability marking (TAPM) traceback scheme. TAPM scheme's main contribution are as follows:

(1) A trust-aware probability marking (TAPM) traceback scheme is proposed in this paper. In TAPM scheme, the marking probability for data packets are adjusted based on the trust of source node. Whereas the marking probability is high for low trust nodes, the malicious nodes can be located quickly. The marking information of most trustable nodes is low, so the amount of data transmitted to sink is less, the network lifetime can be improved. The less locate time for malicious nodes can ensure the network security.

(2) This paper puts forward the active detection traceback scheme, which attempt to speed up locating malicious node. In previous schemes, the system can only get location information from the marking tuples which is generated by malicious nodes, so it is a passive defense approach. The time for locating malicious nodes is long and uncertain. The second innovation for TAPM scheme is active detection, nodes send lightweight detect packets to sink along malicious nodes, which greatly enhance the effectiveness of TAPM scheme.

(3) TAPM scheme has good adaptability and extensibility. In previous schemes, the marking probability of the nodes are pre-determined, it can't change when the network is attacked, so the system can spend more time and cost to determine the position of malicious nodes, and the efficiency for traceback is low. But in TAPM scheme, when a node is regarded as a suspect node by the system, the detection packets can be sent for obtaining more marking information of malicious nodes, and the marking probability of the node will be increased. So more marking information about malicious nodes can be obtained. Thus it has good efficiency and adaptability.

(4) Through our extensive theoretical analysis and simulation study, it shows that: compare to the probability marking (PM) scheme with a probability of 0.8, TAPM scheme can reduce the total number of marking by 67.16% and improve network lifetime by 12.99%-36.61%.The traceback time is only 1/2-1/10 than that of PM scheme.

The rest of this paper is organized as follows: In Section 2, the related works are reviewed. The system model and problem statement are described in Section 3. In section 4, a novel TAPM scheme is presented. Security performance analysis is provided in section 5. Experimental result and comparison is conducted in section 6. We conclude in section 7.

2 Related work

Tracking technology (traceback technology) was first used in the IP network [17-19], the goal is to collect the routing information in the network, and then the system rebuild the network topology and locate the position of malicious nodes, so as to make the system remove malicious source to ensure network security [17-19].

But there are much difference between WSNs and cable IP network, some effective traceback schemes in IP network are not suitable for sensor network. Mobile Ad hoc Networks (MANETs) have emerged as a topical research area in recent years, propose a cross-domain Session Initiation Protocol (SIP) to scale across domains and deal with outbound requests using the reputation method [20, 21]. However, some modified traceback scheme can suitable for sensor network. There are mainly two schemes apply to the traceback in WSNs. One is packets marking [22], another is logging scheme [12]. Packets marking is an effective and popular traceback scheme in the wireless sensor network (WSN). In packets marking scheme, when sensor nodes forwarding packets, nodes add his ID and other information (i. e marking tuples) to packets. After Sink receives the data packets, Sink node reconstruct the path to the source nodes by reading marking tuples. If the source node is malicious node, the system will block or isolate the malicious node. The advantages of this scheme are: it is a simple protocol and almost no storage space requirement for nodes. The shortages are: the number of marking tuples will growth with the packets forwarding to Sink, but the data packets

must to be divided into many pieces in order to send the data packets to Sink, which not only increase the conflict of routing, but also damage the network lifetime [9-11].

The probability marking (PM) scheme is proposed in Ref. [10], in PM scheme, nodes mark data packets with a certain probability. Obviously, in PM scheme, if the specified marking probability is p, the proportion between the amount of marking and the amount of reduced marking in PM scheme is $(1-p)$. Thus, PM scheme can effectively improve the network lifetime. In order to further reduce the amount of marking information, the following researches put forward the scheme that each packet can be marked at most k times [10]. In such scheme, every node marks packet with a certain probability. But after the data packet is marked *k* times, the following nodes are to replace the information that has been marked, which can ensure each packet can be marked at most k times, this scheme is further reduce the amount of marking, so as to further improve the network lifetime. But the shortages of this scheme are: because each packet can be marked at most k times, when the Sink collects the same data packets, the received marking information is far less than packets marking, so the Sink can rebuild one path to the source node when Sink receives more data packets, thus the traceback time is big. If every packet is limited to be marked k times at most, nodes near to the Sink is to have higher probability to be marked. To overcome this deficiency, Ref. [10] presented a fair probability distribution mark scheme, which has a good effect. Ref. [23] proposed a hybrid method for tracking mobile objects with high accuracy and low computational cost.

The scheme based on log (logging) is another kind of tracking technology for malicious nodes [1]. In this scheme, when the marking field in data packets is large, the marking information is stored in the nodes' memory, then the nodes forwarding packets with unloading marking information to next node. In the process of traceback, Sink rebuilds a path from Sink to the source node through querying stored marking information in those nodes. In the traceback scheme based on logging, the proportion of mark information in packets is relatively low. Most of marking information is stored in sensor nodes. Once victims perceive the attack, or need a traceback, inquiry request is sent to required nodes. This information reserved in logging is sent to the Sink to reduce the amount of data received by the sink. So the advantage of this scheme is to have high network lifetime, Logging scheme has the shortage that node requires large storage capacity to store mark information, especially the area near to the Sink which has higher marking probability than the area far from the Sink, thus it requires large storage capacity.

The CPMLT(combined packet marking and logging scheme for traceback, CPMLT) scheme has been proposed in Ref. [12], which combine marking and logging. In CPMLT scheme, a data packet can be marked at most k times, each node marking data packets with a certain probability, nodes logging data packets after been marked k times. It is a compromise in the node's storage capacity and the network life.

A Logging joint Marking (LM) traceback scheme is proposed in Ref. [13]. Compare to the previous schemes, the most important improvement are: in the previous traceback schemes, the energy consumption and storage space of the area near to the Sink are seriously insufficient, much storage capacity and energy left in area far from Sink area. In LM scheme, packets are also marked at most k times, each packet starts logging after been marked k times. When the node's storage space is not enough, data packets will be migrated to area far from the Sink, which can improve the network lifetime. So the scheme can make full use of the residual energy and storage space.

3 The system model and Problem statement

3.1 The System Model

(1) We consider a WSN consisting of m homogenous static sensor nodes $v_i | i \in \{1..m\}$ and the sink node is

v_0, $\mathcal{M} \triangleq \{v = v_0, v_1, v_2, \ldots, v_m\}$ deployed over a 2-D round surveillance field, the network radius is R. Sink node v_0 is the center of the network. The communication radius of sensor nodes is r, the energy of Sink node is unlimited. Sensor nodes monitor their surroundings and once an event happened, nodes report to the Sink through multi-hop [24-26].

(2) The wireless sensor network is deployed in a hostile environment. We consider the following attack scenario. One compromised node is used to launch a false data injection attack to exhaust the network resources, designated as attack node or source node [11]. In case of an attack, the nodes marking passed packets with a certain probability $\mathcal{P}_i$, the system can determine malicious source through those marking information which is similar to cyber forensics technologies. The malicious nodes are considered as a small proportion of the network [1, 10].

(3) The marking probability of the nodes are specified by the system (such as by Sink through the broadcast, or by controlling information). The nodes mark data packets with the specified marking probability. The nodes will expose themselves if they don't mark data packets with the specified marking probability and the attack goal will not be achieved [27]. Therefore, the nodes are considered to abide by the given probability to marking packets.

3. 2 Energy consumption model

$$E_t = lE_{elec} + l\varepsilon_{fs}d^2, \quad if \quad d < d_0$$
$$E_t = lE_{elec} + l\varepsilon_{amp}d^4, \quad if \quad d > d_0 \tag{1}$$

$$E_r = lE_{elec} \tag{2}$$

The energy consumption model adopted in this paper is same as Ref. [7, 13], the energy consumption for transmission E_t denote in Eq. (1), and energy consumption E_r for receiving denote in Eq. (2). E_{elec} represents transmitting circuit loss. Both the free space (d^2 power loss) and the multi-path fading (d^4 power loss) channel models are used in the model, depending on the distance between the transmitter and receiver. ε_{fs} and ε_{amp} are respectively the energy required by power amplification in the two models. l denotes the data bits. The above parameter settings can be seen in [7, 13]. The above parameter settings are given in *Table 1*, as adopted by Ref. [7, 13]

Table 1 network parameters

Parameter	Value
Threshold distance (d_0) (m)	87
Sensing range r_s (m)	15
E_{elec} (nJ/bit)	50
e_{fs} (pJ/bit/m^2)	10
e_{amp} (pJ/bit/m^4)	0.0013
Initial energy (J)	0.5

3.3 Problem statement

The main focus of this paper is to design a new effective TAPM scheme to traceback the DoS/DDoS attacks in WSNs. The goal of the TAPM scheme is to locate the malicious source as soon as possible at less cost, which can be categorized in following aspects:

(1)Maximization of network lifetime. The basic goal of application requirement is to maximize network lifetime. The network lifetime can be defined as the time of the first node that dies [7, 13]. Since after the first node dies, it may affect the connectivity and coverage of the network severely leading the network cannot play a proper role.

Hence, the definition of network lifetime in this paper is consistent with references [7, 13], which is defined as the time elapsed until the first sensor node in the network depletes its energy. We denote E_i as the energy consumption of node v_i in one round. E_{init} is the initial energy of node $\boldsymbol{v_i}$. The formula of maximizing network lifetime can be expressed as follows:

$$\max(\Gamma) = \max_{i\in\{1..m\}} min\ (E_{init}/E_i) \tag{3}$$

(2) Scheme can locate attack source quickly when defending against attracts.

The locate attack malicious source time $\mathcal{T}$ is evaluated in terms of the number of marking information for the attack paths can be reconstructed. Obviously, in process of reconstructing the attack paths, if traceback scheme marks more data packets, then the system can collect much more marking information shortly and can locate malicious node quickly, , and then revoke compromised nodes in order to ensure the confidentiality of data traversing in the network [28, 29]. Therefore, $\min(\mathcal{T})$ means to maximize marking information. $\mathscr{b}_i$ denotes the number of marking information by node v_i in a unit time, and so

$$\min(\mathcal{T}) = \max_{i\in\{0..m\}} \sum \mathscr{b}_i \tag{4}$$

Generally, compromising optimization exists in the performance indexes above. In summary, the optimization purpose of the scheme in this paper is

$$\begin{cases} \max(\Gamma) = \max\limits_{i\in\{1..m\}} min\ (E_{init}/E_i) \\ \min(\mathcal{T}) = \max\limits_{i\in\{0..m\}} \sum \mathscr{b}_i \end{cases} \tag{5}$$

4 Trust-aware probability marking traceback scheme design

4.1 Research motivation

In this paper, the motivation is based on the following two aspects:

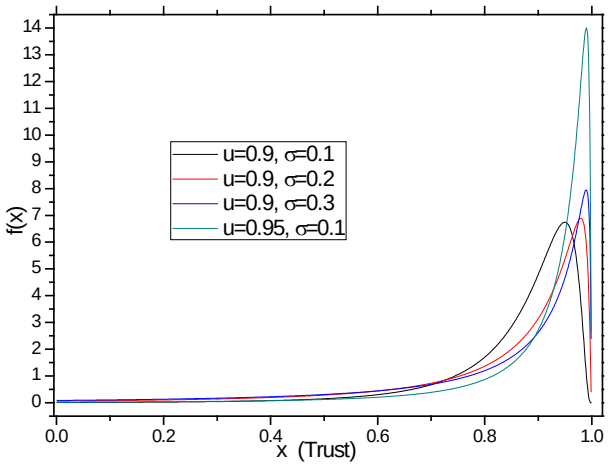


Fig. 1 probability density function of the nodes trust

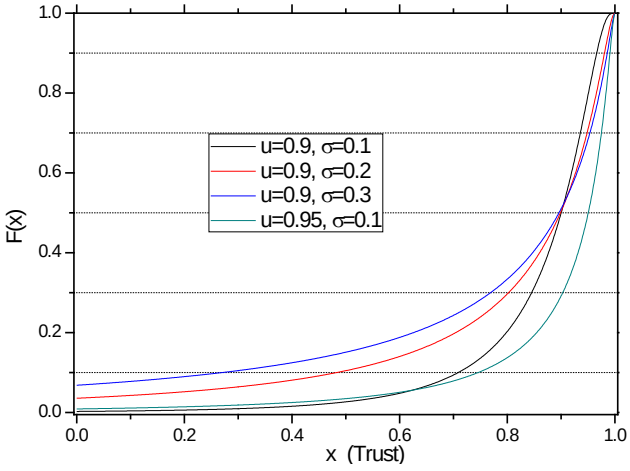


Fig. 2 distribution function of the nodes trust

(1) In previous probability marking (PM) schemes, every node v_i marks each passed packet with equal marking probability, the marking probability for data packet generated by malicious nodes is the same. Due to most nodes in the network are "good" nodes, it is important to decrease marking probability of packets which are generated by "good" nodes, and increase marking probability of packets which are generated by malicious nodes. So the network

lifetime can be improved and the traceback ability for malicious nodes can be enhanced.

The ratio of malicious nodes to all nodes in the network is relatively small, generally less than 10%. The probability density function of node trust can be described as logarithmic normal distribution [30], as can be seen in Fig. 1. Most nodes in the network are credible, so trust of most nodes are about 0.9. It can be seen form Fig. 2 that the proportion of node trust<0.3 is about 10%, which is in line with the actual situation that the proportion of malicious nodes in the network is small. For node trust distribution, it is described by the function of logarithmic normal distribution, but the other functions can also be used to describe the credibility distribution. As long as the function reflects that most nodes are "good" nodes in the network, it doesn't influence the conclusions.

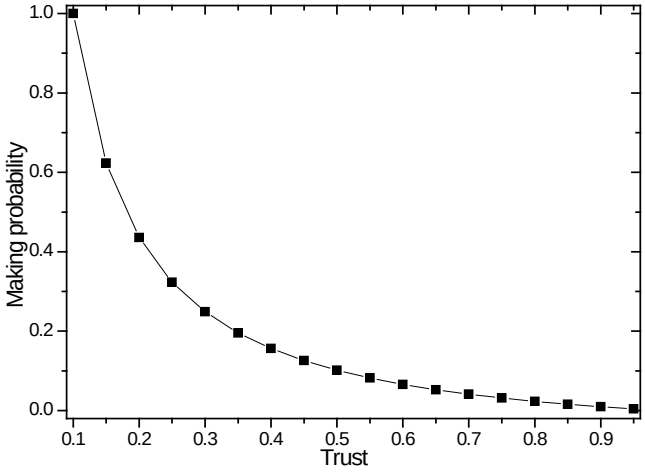


Fig. 3 marking probability according trust

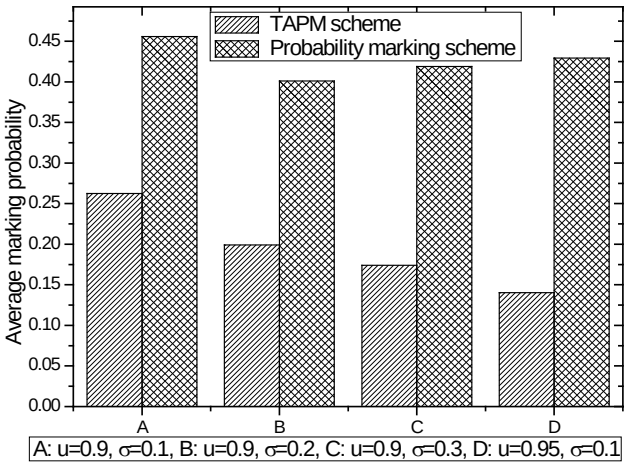


Fig. 4 The average marking probability

Based on marking scheme of TAPM scheme, a marking method with non-linear relationship between marking probability and nodes' credibility is proposed, as can be seen in Fig. 3. In TAPM scheme, the marking probability rise sharply with the increase of node trust when node's credibility is low, the marking probability is low when the node's credibility is high, which can meet the design goal for TAPM scheme. Fig. 4 shows the weighted average marking probability under different scheme. Though the marking probability for nodes with low credibility are more than 80% in TAPM scheme, the average marking information is only about 20% (see from Fig. 4). Even the marking probability is more than 40% in PM scheme, the effect is not as good as TAPM scheme, which shows that the traceback ability for malicious nodes in TAPM scheme is better than that of other schemes. On the other hand, due to decreasing marking probability of "good" nodes in the network, the number of transmission information is decreased, which can improve the network lifetime (see from Fig. 5).

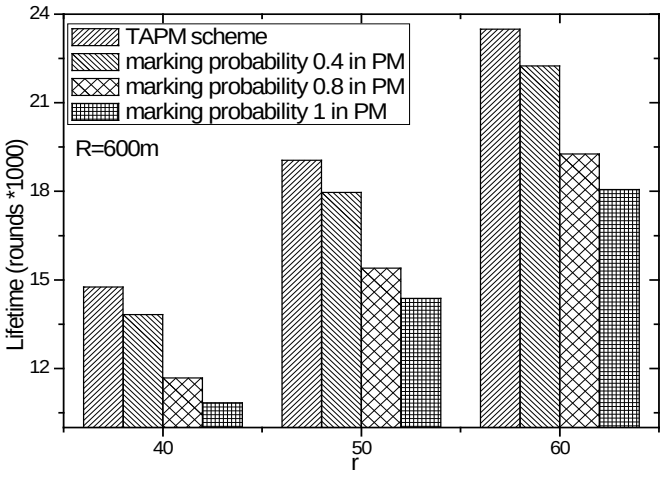


Fig. 5 The lifetime

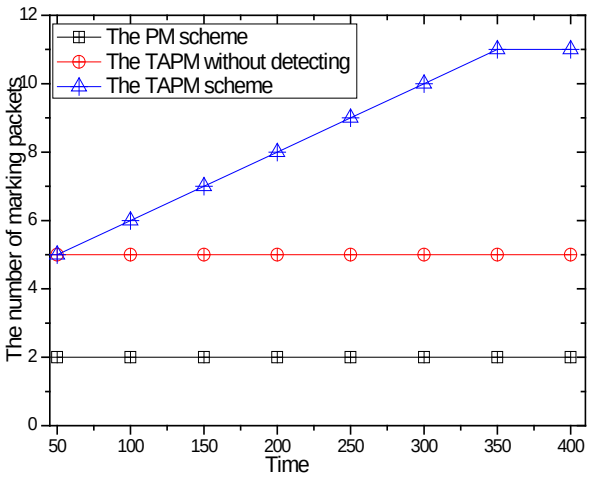
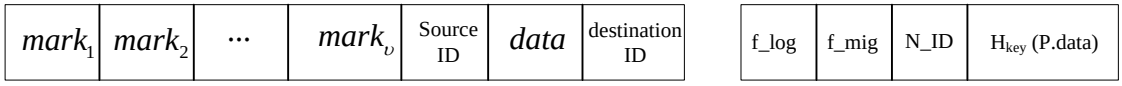


Fig. 6 The number of marking tag fluctuate based on time

(2) To the best of our knowledge, the previous schemes use passive marking strategy, namely, data packets only can be marked when data packets are sent to the Sink. Therefore, this paper propose a kind of active marking probability scheme. In this scheme, for the suspicious nodes, nodes in the upstream of the routing generate probe packets. The difference between probe packet and data packet is that the length of data field is 0. Probe packets can route to the Sink bypass malicious nodes, so that they can increase the marking probability for specific suspicious nodes to locate malicious nodes quickly. Such as in Fig. 6, for suspicious nodes, if the system doesn't adopt active detect strategy, the received marking packets are essentially unchanged per unit time in PM scheme (see from Fig. 6). And if the system doesn't adopt active detect mechanism in TAPM scheme, although the received marking packets per unit time is higher than that of PM scheme, it still can't meet the requirement of application (see from Fig. 6). If TAPM scheme adopts active detect scheme, the received marking packets can increase with the need of traceback, so as to completely get rid of the limitations of the original passive scheme, which make higher performance and more extensive applicability for TAPM scheme.

4.2 The overview of TAPM scheme

The structure of packets in TAPM scheme is illustrated in Fig. 7 which is similar Ref. [13]. Each packet is mainly composed of the following field: (1) Data field, represent the comment of data packets; (2) Source ID, denote the source node ID of packet; (3) Destination ID, denote the destination node ID of packet; (4) Marking field, represent the marking information which is composed of multiple marking tuple. However, each marking tuple includes multiple fields, such as in Fig. 7(b). It mainly includes: ① N-ID, denote the node ID of node which create marking tuple; ②f-log, the marking for logging, it is used in traceback scheme with marking and logging method. If f-log=1, it denotes marking information of data packets are stored in this node, or f-log=0; ③ f-mig, denote whether marking information are migrated to other nodes which is adopted in Ref. [13]; ④ Hkey (P.data), denote the value of data after hash data in data packets, it is used to test the consistency of data packets.



(a) The structure of packets (b) The structure of mark tuple

Fig. 7 The structure of packets

The overview of trust-aware probability marking (TAPM) traceback scheme can be illustrated in Fig. 8. In TAPM scheme, sensor nodes mark each data packets with a certain probability based on source nodes' trust. If one node marks data packets, it can form marking tuple in Fig. 7(b), then add this marking information to marking field and forward to next node (or Sink). In case of an attack, the system can reconstruct attack path by those marking information to locate malicious source.

The difference between TAPM scheme and previous schemes are:

(1) Determine marking probability. When a data packet arrives to a node, firstly, the node reads source ID field of the data packet to determine the source node ID of the packet. Secondly, according to the credibility of source node, impose high marking probability of the packets which generated by suspicious nodes, and low marking probability of the packets which generated by trusted nodes according to node trust in section 4.4.

(2) Generate and marking of detect packets. The second difference between TAPM scheme and previous schemes is: TAPM scheme actively generates probe packets with some marking information to collect enough marking information for locate malicious nodes quickly. It is shown in Fig. 8. Considering that node v_5, v_{11}, v_{13} are suspicious malicious nodes according to the received marking information, in order to improving the marking probability of the generated packets by suspicious malicious nodes, some nodes are selected in the routing upstream area of suspicious malicious nodes to send probe packets to Sink to locate malicious nodes quickly. In Fig. 8, the system chose three area I, II, III to send probe packets, the nodes send a certain number of probe packets to the Sink

in these areas, the nodes which the probe packets passed by mark the probe packets with marking probability 1, thereby can locate malicious nodes quickly.

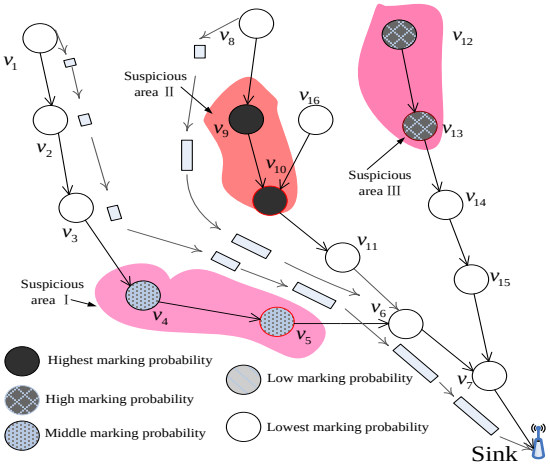


Fig. 8 Illustrate of trust-aware probability marking (TAPM) scheme

The overview for TAPM scheme is discussed in this section. The two key issues for TAPM scheme are: (1) how to determine its marking probability based on nodes' credibility. (2) How to choose nodes to send probe packets, how many probe packets should be sent?

4.3 Computing of trust and marking probability

The Sink collects marking information from sensor nodes, then evaluate the trust of sensor nodes. This paper assumes that the network have the same data detection in intrusion detection system [1], which can detect each packet, then obtain each node's credibility. Because this paper focuses on how to construct an effective method for obtaining the marking information of sensor nodes to locate malicious nodes, the method of trust evaluation for nodes is same as Ref. [31] which is used in the application layer to evaluate and manage trust. So we assume that the trust of sensor nodes can be get. More details can be found in [1, 31]. Considering the credibility for node j in time t_i is $c_{j,t_i}|\ 0 \le c_{j,t_i} \le 1$ when adopt trust evaluation method of Ref. [31]. The value of credibility c_{j,t_i} is closer to zero, the lower its credibility is, and the bigger the probability of malicious nodes is. On the other hand, the value of c_{j,t_i} is closer to 1, which show that nodes are trusted. The system can form multiple evaluation results if the system receives data packets with a marker of node j many times in a period of time. The evaluation result in the K_{th} time for node j is: $c_{j,t_i}^{(k)}$. The series of evaluation results can be expressed as: $c_j=\left\{c_{j,t_1}^{(1)}, c_{j,t_2}^{(2)}, \ldots c_{j,t_w}^{(w)}\right\}$, where $0 \le c_{j,t_k}^{(k)} \le 1, k \in [1..w]$, w is the biggest number of effective trust evaluation. The elements in $\left\{c_{j,t_1}^{(1)}, c_{j,t_2}^{(2)}, \ldots c_{j,t_w}^{(w)}\right\}$ are in interaction time order. $c_{j,t_1}^{(1)}$ is the longest evaluation result from now and $c_{j,t_w}^{(w)}$ is the last evaluation result. The total trust evaluation results of node j are as follows:

$$c_j = \begin{cases} \sum_{k=1}^{w} c_{j,t_k}^{(k)} \cdot \hbar(k)/w, & w \neq 0 \\ 1, & w = 0 \end{cases} \quad (6)$$

$\hbar(k) \in [0,1]$ is attenuation function, it is used to make reasonable weighting of the trust evaluation at different times. The result of trust evaluation in the time closer to the current time have more conducive to judge the credibility of nodes. More weight should be given to new result for trust evaluation, thus the attenuation function is defined as:

$$\hbar(k)=\begin{cases}1, & k=w\\ \hbar(k-1)=\hbar(k)-1/w, & 1\le k\le w\end{cases} \tag{7}$$

The value c_j of trust evaluation for any node v_j can be obtained by the system. After the system obtains the trust of each node, it can determine the marking probability function according the results. The marking probability function should have the following property.

Character 1: The higher nodes' credibility is, the lower the marking probability is. The marking probability should be increase with the decrease of nodes' credibility. However, the increase ratio of marking probability function should be bigger than the decline rate of its trust.

The marking probability function with those characteristics can ensure that the marking probability is high when node trust is low which is beneficial to locate malicious nodes quickly. But the marking probability of nodes with high credibility is low, which can improve the network lifetime. Through the above analysis, it is easy to choose appropriate marking probability function. The function is available if it meets the character 1, so it is convenient to design the marking probability function. For example: sina function is a function that the change quantity of the value in this function is less with the change of independent variable in the first, and then the change quantity is more, which can meet the requirement of character 1. Therefore, this paper constructs the marking probability function as Eq. (8) to illustrate the effectiveness of TAPM scheme. It is important to note that Eq. (8) is one of marking probability function meets the character 1. There are many functions which can meet the **character 1**. But this doesn't influence the conclusions of this paper.

$$\mathcal{L}(c)=\left(\frac{1}{\sin(c)}-\frac{1}{\sin(\mathfrak{b})}\right)\Big/\left(\frac{1}{\sin(\mathfrak{a})}-\frac{1}{\sin(\mathfrak{b})}\right),\quad c\in[\mathfrak{a},\mathfrak{b}] \tag{8}$$

Where c is the trust of node, which is denoted in Eq. (7) for node j. $\mathfrak{a}$ and $\mathfrak{b}$ are two constant, respectively denote the upper and lower bounds value of node trust. Fig. 3 shows a specific example of Eq. (8).

4.4 The number of detect packets

The other important innovation of TAPM scheme is to generate detection packets. Thus this section discusses how to determine the number of detection packets, and decides which nodes need to generate detection packets. The number of detection packets need to be generated depends on the following two factors: (1) It is better to protect cyber security for generating more detect packets. ① The number of detection packets are related to node's credibility. Obviously, the lower node trust is, the more number of detection packets need to be generated. ② The time allowed for locating malicious nodes. If the system need to locate malicious nodes as fast as possible, it needs to generate many detection packets as soon as possible. (2) But large number of detect packets can harm network lifetime, it should reduce the number of detection packets in order not to reduce the network lifetime. Therefore, the number of detection packets need to take into account these two factors, so as to obtain an optimal equilibrium point, which can safeguard cyber security and ensure higher network lifetime.

Like many systems [32], the number of detection packets in TAPM scheme exist equilibrium point between the cost and obtained "payoff". Generally speaking, in the early stages, the number of detection packets is small, increasing the number of detection packets a little can greatly shorten the time to locate malicious nodes, so as to take corresponding measures to remove malicious nodes. After removing malicious nodes, the system does not need generate so much detection packets, and the marking probability of nodes can be reduced. To do so, the system can locate malicious nodes quickly with higher security and decrease the amount of marking information at low cost. But when the number of detection packets arrive to a certain amount, the increase of detection packets can bring a little increase in payoff. Therefore, it is not a wise decision to increase the number of detect packets in this time. In this paper, some conclusions can be obtained from many researches: the utility function between obtained payoff by

system and obtained sent detection packets (cost) is one of non-linear functions [32], it meets the practical of TAPM scheme. Namely: when the number of detection packets are less, payoff is rising quickly with the growth of detection packets, so its utility is high. But after arrive to a certain degree, the number of generating detect packets increase, the growth of obtained payoff is very small, so its utility is small. The same as Ref. [1, 32], Eq. (9) is adopted as utility function for TAPM.

$$\mathcal{G}(\mathcal{D}) = \beta \log(1 + \mathcal{D}) + \gamma\mathcal{D} \quad (9)$$

Where β, γ are constant parameters. The obtained payoff is the difference between utility function and cost. The system pays the cost for energy consumption for generating detection packets. The cost can be calculated based on energy consumption function Eq. (1) and the number of generating data function Eq. (2). The cost paid is a linear relationship with the number of sending detect packets, that is, $\mathfrak{E}\mathcal{D}, \mathfrak{E}$ are constant coefficient. The payoff function can be shown:

$$\mathfrak{U} = \beta \log(1 + \mathcal{D}) + \gamma\mathcal{D} - \mathfrak{E}\mathcal{D} \quad (10)$$

Theorem 1. In TAPM scheme, the optimal number of detect packets $\mathcal{D}$ to maximize its payoff is given by Eq. (11):

$$\mathcal{D}^* = \frac{\beta}{\mathfrak{E} - \gamma} - 1 \quad (11)$$

Proof: Obviously, $\mathcal{D}$ is a bounded closed set in Euclidean space, and the payoff function Eq. (10) is continuous on its scheme space [32]. The payoff function is concave function, it is proved in the following formula, and there is a Nash equilibrium point. The first and second order derivatives of Eq. (20) with respect to $\mathcal{D}$ are:

$$\frac{\partial \mathfrak{U}}{\partial(\mathcal{D})} = \frac{\beta}{1 + \mathcal{D}} + \gamma - \mathfrak{E}$$

$$\frac{\partial^2 \mathfrak{U}}{\partial^2(\mathcal{D})} = -\frac{\beta}{(1 + \mathcal{D})^2} < 0$$

Since $\frac{\partial^2 \mathfrak{U}}{\partial^2(\mathcal{D})} < 0$, $\mathfrak{U}$ is strictly concave in $\mathcal{D}$. Hence, the optimal $\mathcal{D}$ that maximizes $\mathfrak{U}$ is determined by letting the marginal utility $\frac{\partial \mathfrak{U}}{\partial(\mathcal{D})}$ equal to 0, i. e. $\mathfrak{E} - \gamma = \frac{\beta}{1+\mathcal{D}}$. $\mathcal{D} = \frac{\beta}{\mathfrak{E} - \gamma} - 1$, which leads to Eq. (11).

■

The system can determine the number of nodes after Theorem 1 determine the number of optimized detect packets, the number of generating detect packets by one nodes are $\mathfrak{T}$ in a detection cycle, and the number of requirement detection packets are $\mathcal{D}$ in a detect cycle. So the number of selected nodes which is used to send detect packets are:

$$\mathbb{v} = \mathcal{D}/\mathfrak{T} \quad (12)$$

After obtain $\mathbb{v}$ detection packets which need to be generate, it is easy to choose nodes to generate detect packets. The method is that select $\mathbb{v}$ nodes which nearest to malicious nodes and their packets route to Sink through malicious nodes. For example, in Fig. 1, considering the node v_{10} is suspicious node which needs to generate detection packets. it can calculate $\mathbb{v}$=3 according Eq. (12). It should select 3 nodes v_9, v_8, v_{16} nearest to node v_{10} to generate detect packets. If $\mathbb{v}$=2, nodes v_9, v_{16} should be selected, because the distance between node v_9 or v_{16} and node v_{10} is shortest.

4.5 TAPM traceback algorithm

Through the above analysis, the algorithm about TAPM scheme is given. Algorithm has 2 executive body, (1) the first is the Sink. The task for the Sink are: (a) Trust evaluation, in order to obtain trust degree of each node; (b) calculate the marking probability of nodes using Eq. (8); (c) Calculate the number of sent probe packets of suspicious nodes whose credibility below a certain threshold according to Eq. (11); (d) Calculate the number of nodes for sending detect packets using Eq. (12); (e) Select $\mathbb{v}$ nodes to send probe packets, on the basis of the principle of

closest to the suspicious nodes. (f) Report the system's decision to related nodes, which include the marking probability of each node, the number of nodes for sending detect packets, and the number of sent probe packets. (2) The second executive body is sensor node, it marks each passed packet with a certain probability according to the given probability. If it is a detection node, it sends detect packets according to the given time and send rate. Obviously, the marking probability for probe packet is 1. The algorithm for executive body the Sink and sensor node are shown respectively in algorithm 1 and 2 algorithm.

Algorithm 1. The pseudo-code of TAPM traceback algorithm for Sink

Algorithm 1: The trust-aware probability marking traceback algorithm for Sink

For each time t **Do**
  For each node v_i
    evaluation $c_{i,t}$ of node v_i as Ref. [31];
    computing the trust of c_i node v_i using Eq. (6);
    computing the marking probability $\mathfrak{p}_i = \mathcal{L}(c_i)$ of node v_i using Eq. (8);
    If $\mathfrak{p}_i \le \Delta$ then //Δ is the trust threshold
      computing the optimization number of detect packets $\mathcal{D}_i^*$ using Eq. (11);
      computing the $\mathbb{v}$ using Eq. (12);
      select the candidate node set to send detect packets based on the nearest principles, which is
        $\mathbb{D}_i = \{ v_j \} \ | v_j \in$ the top $\mathcal{D}_i^*$nearest node of node v_i
      the detect packets z_j of each detect node for node v_i is $\mathbb{v}/\mathbb{D}_i$;
    End if;
    notice each node with $v_i's$ information include $\{\mathfrak{p}_i\}$;
    notice $v_j \in \mathbb{D}_i$ with z_j;
  End for
End For

Algorithm 2. The pseudo-code of TAPM traceback algorithm for Sensor node

Algorithm 2: The trust-aware probability marking traceback algorithm for sensor node

Input: node v_i receive the number of detect packets z_i, each nodes' marking probability $\mathfrak{p}_k$;
Output: Forward the packet to next hop node

For each receive packets $\mathbb{p}$ **Do**
  If $\mathbb{p}$ ∈data packets then // $\mathbb{p}$ is not detect packets
    v_s =Get_source_node($\mathbb{p}$); // v_s is the source node of packets $\mathbb{p}$
    marking packets with $\mathfrak{p}_s$;
    If marking this packet $\mathbb{p}$ then //with a certain probability for marking packets $\mathbb{p}$
      $\mathbb{p}$=Ekey(|f_log|f_mig|N_ID|Hkey ($\mathbb{p}$.data)| $\mathbb{p}$); //Ekey stand data encryption,
    End if // Hkey denote hash
  Else // $\mathbb{p}$ is detect packets
    $\mathbb{p}$=Ekey(|f_log|f_mig|N_ID|Hkey ($\mathbb{p}$.data)| $\mathbb{p}$); //detect packet marking probability is 1
  End if
  forwarding packet $\mathbb{p}$ to next node; //Let processed packet $\mathbb{p}$ forwarding to next node
End for
For $z = 1$ to z_i Do
  Produce packet $\mathbb{p}$ which a data field is null;
  $\mathbb{p}$=Ekey(|f_log|f_mig|N_ID|Hkey ($\mathbb{p}$.data)| $\mathbb{p}$); //marking this packet;

16: forwarding packet $\mathbb{p}$ to next node;

17: End For

The computation complexity of the TAPM traceback algorithm could be discussed as follow: algorithm 1 and algorithm 2 are easy. Algorithm 1 is performed by the sink. The sink calculates node trust, then gives different marking probability according to node trust. So its computation complexity is $O(kn)$, where k is a constant, n is the number of nodes in the network. For algorithm 2, each node calculates hash function and encryption function. Its computation complexity is $O(\mathfrak{H}(lengt(data) + \mathfrak{E}(lengt(data))$, where $\mathfrak{H}(m)$ is the requirement time function for hash data packet with m bits, and $\mathfrak{E}(m)$ is the requirement time for encrypt data packet with m bits. Because the length of data packets in the network are short, its computation complexity is relatively small.

5 Performance Analysis of TAPM scheme

5.1 The number of data and the network life for node

This paper adopts logarithmic normal distribution to describe distribution of node trust [30], because logarithmic normal distribution can represent that there are many nodes with high credibility and less malicious nodes in the network. The density function of logarithmic normal distribution is:

$$f(x) = \frac{1}{x\sigma\prime\sqrt{2\pi}} e^{-\frac{(\ln x-\mu\prime)^2}{2(\sigma\prime)^2}}, \quad x > 0 \tag{13}$$

Where parameters μ' and σ' determine the function of logarithmic normal distribution (see from Fig. 1). The average value μ and variance σ^2 in logarithmic normal distribution are respectively:

$$\mu = e^{\mu'+(\sigma\prime)^2}/2, \quad \sigma^2 = e^{2\mu'+(\sigma\prime)^2}\left(e^{(\sigma\prime)^2} - 1\right) \tag{14}$$

The probability rule for the random variable in the logarithmic normal distribution is: the value closer to μ, the higher the probability is. The smaller the σ is, the more the distribution concentrate on the area near to μ, the bigger the σ is, the more the distribution dispersed. Therefore, we can get the value of $\mu'$and σ' if node's μ and σ have to be known.

$$\mu' = \ln\mu - \frac{(\sigma\prime)^2}{2}, \quad \sigma' = \sqrt{\ln\left[\left(\frac{\sigma}{\mu}\right)^2 + 1\right]} \tag{15}$$

Analyse the number of data and marking information under the condition of knowing the distribution function of node trust. Considering the network radius is R, the transmitting radius is r, the event generation rate of the network is λ, each node sends data packets to the Sink through the shortest routing method. Ref. [29] has proved that the number of sending data packets of the nodes at $l = hr + x$ distance from the Sink are:

$$d_l = \left((z+1) + \frac{z(z+1)r}{2l}\right)\lambda \mid z = \left\lfloor\frac{R-l}{r}\right\rfloor \tag{16}$$

Theorem 2. In TAPM scheme, the number of marking tuples for marking passed data packets by nodes at $l = hr + \mathcal{y}$ distance from Sink are:

$$\mathfrak{m}_l = \left((z+1) + \frac{z(z+1)r}{2l}\right)\lambda \int_a^b \mathcal{L}(x) f(x) \tag{17}$$

Proof: According to Eq. (16), the number of forwarded data packets by nodes at l away from Sink are d_l. Because

data packets are come from the network nodes, the marking probability of the packets are determined based on the nodes trust. Therefore the marking probability also obey the distribution of node trust. Considering the distribution density function of node trust is $f(x)$, and the transformation function from node trust to marking probability is $\mathcal{L}(x)$. The expected marking probability for data packets can be obtained:

$$\mathbb{E}_{\mathfrak{f}} = \int_a^b \mathcal{L}(x) f(x) \tag{18}$$

Where $\mathbb{E}_{\mathfrak{f}}$ said the expected marking probability in node trust distribution function $\mathfrak{f}$, the number of assumed data packets are d_l, so the number of marking tuple for marking data packets are:

$$\mathfrak{m}_l = d_l \mathbb{E}_{\mathfrak{f}} = \left((z+1) + \frac{z(z+1)r}{2l}\right) \lambda \int_a^b \mathcal{L}(x) f(x) \tag{19}$$

■

In theorem 2, the number of marking tuples for the nodes at l away from Sink are $\mathfrak{m}_l$ in the condition of node trust distribution function $\mathfrak{f}$. When the distribution function $\mathfrak{f}$ of node trust is logarithmic normal distribution, the transformation function from node trust to marking probability is used by Eq. (8). The number of marked marking tuple by nodes at l away from Sink are $\mathfrak{m}_l$:

$$\mathfrak{m}_l = \int_a^b \left(\left(\left(\frac{1}{\sin(x)} - \frac{1}{\sin(\mathfrak{b})} \right) / \left(\frac{1}{\sin(\mathfrak{a})} - \frac{1}{\sin(\mathfrak{b})} \right) \right) \frac{1}{x\sigma\prime\sqrt{2\pi}} e^{-\frac{(\ln x - \mu\prime)^2}{2(\sigma\prime)^2}} \right) \left((z+1) + \frac{z(z+1)r}{2l} \right) \lambda \tag{20}$$

Theorem 3. In TAPM scheme, the number of total marking tuple by nodes at $l = hr + \mathcal{y}$ away from Sink are:

$$\mathbb{M}_l = \mathbb{E}_{\mathfrak{f}} \sum_{i=0}^{z} \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor \tag{21}$$

Proof: nodes at l away from Sink forward the marked marking tuples in $l+r,\ l+2r, \ldots, l+zr$ area. Because $z = \left\lfloor \frac{R-l}{r} \right\rfloor$, d_{l+ir} can be written:

$$d_{l+ir} = \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \tag{22}$$

The number of marking tuples are:

$$\mathbb{M}_l = \mathbb{E}_{\mathfrak{f}} \sum_{i=0}^{z} d_{l+ir} = \mathbb{E}_{\mathfrak{f}} \sum_{i=0}^{z} \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor \tag{23}$$

■

Inference 1: In PM scheme, consider the marking probability is $\mathcal{P}_p$, the ratio of the number of marking tuples of TAPM scheme over PM scheme is:

$$\varphi = \mathcal{P}_p / \mathbb{E}_{\mathfrak{f}} \mid \mathbb{E}_{\mathfrak{f}} = \int_a^b \mathcal{L}(x) f(x) \tag{24}$$

Proof: According to theorem 2, for TAPM scheme, the number of marked marking tuple for passed data packets by nodes at $l = hr + x$ away from Sink are: $\mathfrak{m}_l = \mathbb{E}_{\mathfrak{f}} \left((z+1) + \frac{z(z+1)r}{2l} \right) \lambda$. In PM scheme, the marking probability for each node is $\mathcal{P}_p$, the number of marked marking tuple are:

$$\mathfrak{C}_{\ell}^{p} = \mathcal{P}_p \left((z+1) + \frac{z(z+1)r}{2l} \right) \lambda \tag{25}$$

So: $\varphi = \mathfrak{C}_{\ell}^{p} / \mathfrak{m}_l = \mathcal{P}_p / \mathbb{E}_{\mathfrak{f}}$.

■

For TAPM scheme, nodes also forward the data of probe packets, it can be seen in theorem 4.

Theorem 4. In TAPM scheme, the number of marking tuples for forwarding the detection packets by nodes at $l = hr + \mathcal{y}$ away from Sink are:

$$\mathbb{B}_l = \frac{\lambda}{\kappa}\mathcal{D}^* \int_a^{\Delta} f(x) \sum_{i=0}^{z} \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor \quad (26)$$

Proof: Based on TAPM algorithm, the proportion of suspicious nodes for probe packets is: $\int_a^{\Delta} f(x)$, those suspicious nodes do a detection in κ data collection cycle (refer to the time for each node send a data packet in the network). According to theorem 1, the number of generated detect packets for each node are $\mathcal{D}^*$. The number of marked marking tuples by nodes at l away from Sink are:

$$\mathfrak{B}_l = \frac{\lambda}{\kappa}\mathcal{D}^* \left((z+1) + \frac{z(z+1)r}{2l} \right) \int_a^{\Delta} f(x) \quad (27)$$

Similar to theorem 3, nodes at l away from Sink forward the marked marking tuples for probe packets of suspicious nodes in $l + r,\ l + 2r, \ldots, l + zr$. Thus the total number of assumed marking tuple are:

$$\mathbb{B}_l = \frac{\lambda}{\kappa}\mathcal{D}^* \int_a^{\Delta} f(x) \sum_{i=0}^{z} d_{l+ir} = \frac{\lambda}{\kappa}\mathcal{D}^* \int_a^{\Delta} f(x) \sum_{i=0}^{z} \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor \quad (28)$$

■

Considering the length of data packets and marking tuple is $\mathfrak{h}$ bits and $\mathfrak{x}$ bits respectively, then in TAPM scheme, according to Eq. (16), Eq. (21), and Eq. (26), the number of data packets d_l and marking tuple $\mathbb{M}_l$ of nodes at l away from Sink can be obtained. So the total amount of information are:

$$\mathfrak{D}_l^{\mathfrak{t}} = d_l\mathfrak{h} + \mathbb{M}_l\mathfrak{x} + \mathbb{B}_l\mathfrak{x} \quad (29)$$

For PM scheme, the total amount of assumed information by nodes at l away from the Sink are:

$$\mathfrak{D}_l^{p} = d_l\mathfrak{h} + \mathfrak{G}_l\mathfrak{x} \mid \mathfrak{G}_l = \mathcal{P}_p \sum_{i=0}^{z} d_{l+ir} = \mathbb{E}_{\mathfrak{f}} \sum_{i=0}^{z} \left(\left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) + \frac{\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor \left(\left\lfloor \frac{R-(l+ir)}{r} \right\rfloor + 1 \right) r}{2(l+ir)} \right) \lambda \quad (30)$$

According to Eq. (1), considering energy consumption for sending 1 bit is $e(r)$, where r transmission distance is. So, the lifetime in TAPM can be get as follow:

$$\ell_{tapm} = \frac{E_{init}}{e(r)\mathfrak{D}_{l_0}^{\mathfrak{t}}} = \frac{E_{init}}{e(r)\left(d_{l_0}\mathfrak{h} + \mathbb{M}_{l_0}\mathfrak{x} + \mathbb{B}_{l_0}\mathfrak{x}\right)} \quad \text{Where } l_0 \text{ is the node closest to the Sink node.} \quad (31)$$

$$\ell_{pm} = \frac{E_{init}}{e(r)\mathfrak{D}_{l_0}^{p}} = \frac{E_{init}}{e(r)\left(d_{l_0}\mathfrak{h} + \mathfrak{G}_{l_0}\mathfrak{x}\right)} \quad \text{Where } l_0 \text{ is the node closest to the Sink node.} \quad (32)$$

5.2 Traceback time analysis

Because it is little meaningless to traceback for "good" nodes. Therefore, traceback time is the requirement time to build a complete path from Sink to malicious nodes. Obviously, the collected marking information of the Sink increase with the large number of collected data packets, and the path from Sink to malicious nodes can be reconstructed. It can be seen that the faster the speed of collecting data packets is in PM scheme without active detection, the shorter the traceback time is. In order to compare to different traceback schemes, therefore, in this paper, we define the requirement time for collecting a date packet of each node as a unit of time, or called a round, the requirement rounds for reconstruct the path from Sink to malicious nodes in traceback time are:

(1) Firstly, analyze the making probability of malicious nodes in TAPM scheme. In TAPM scheme, it adopts

higher marking probability for nodes with low credibility, so the marking probability of most malicious nodes will be very high, but the system cannot always identify malicious nodes accurately, the reason is that: a small number of malicious nodes with low marking probability. In this way, the distribution density function of making probability can also be described by the function of logarithmic normal distribution in Eq. (33), y in Eq. (34) saids the making probability. The distribution function of making probability of malicious nodes is shown in Fig. 9, μ and σ in the distribution function are respectively denote the average making probability and variance of malicious node. So the average making probability is mostly distributed between 0.8 and 1.0 according to the distribution function of malicious nodes.

$$f(y)=\frac{1}{y\sigma\prime\sqrt{2\pi}}e^{-\frac{(\ln y-\mu\prime)^2}{2(\sigma\prime)^2}},\quad x>0 \tag{33}$$

(2) Secondly, analyze trackback time. Set the hops that one nodes away from Sink are $\hbar$, the marking probability of each node is $\wp$, when the node sends n data packets, the probability for the attack paths can be reconstructed is:

$$\mathbb{P}_{tra}=(1-\bar{\wp}^{n})^{\hbar} \tag{34}$$

According Eq. (34), Fig. 10 gives the probability for reconstructing one attack path when a node generates n data packets and $\hbar$=5. When the marking probability is low, $\wp$=0.1, and a node generates 15 data packets, the probability for reconstructing attack path is only 31.58%, and when marking probability $\wp$=0.8, a node only sends 4 data packets, the probability is greater than 99%. When a node generates 9 data packets, the probability is almost 100%. So it is important to improve marking probability of malicious nodes to reduce traceback time. On the contrary, if the probability for reconstructing attack path is $\mathbb{P}_{tra}$, the number of packets need to be generated:

$$n=\frac{\log\left(1-\sqrt[\hbar]{\mathbb{P}_{tra}}\right)}{\log(1-\wp)} \tag{35}$$

(3) The effect of detection in TAPM scheme on traceback time. Eq. (34) and Eq. (35) show that the traceback time is a kind of probability function if the marking probability isn't 1. In TAPM scheme, the marking probability for generate detect packets is almost 1, thus it can ensure that traceback time in TAPM scheme has a certain upper bound, and it is an important improvement.

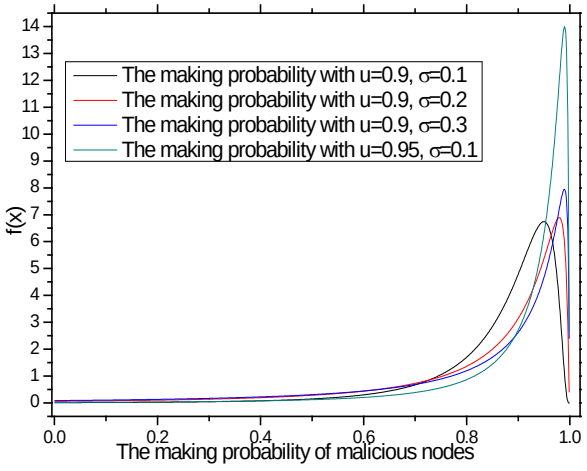


Fig. 9 distribution function of making probability of malicious nodes

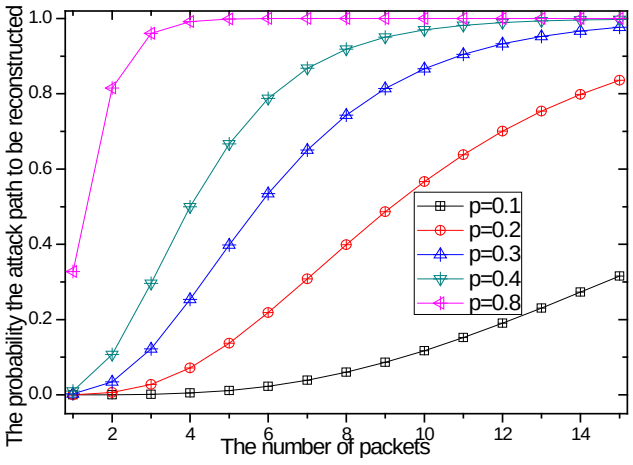


Fig. 10 The relation between the requirement packets and the probability the attack path to be reconstructed ($\hbar$=5)

6 Experimental results

Network parameters are set as follows: the network radius R=500m, r=60m, there are 800 nodes in network, 10% of nodes are malicious nodes, and the system proper adjust the value of the parameters according to the circumstances of the experiments, under normal circumstances, the proportion for malicious nodes are about 5-15%. The packet length are 100 bits, and the length of marking tuple are 10 bits. The requirement time for sending a packet to Sink is called a round. The initial marking probability of nodes is 0.5. The trust of nodes in the network is 0.5.

6. 1 Trackback time

Traceback time is an important performance indicator for cyber security in traceback scheme. So this section analyzes traceback time in TAPM scheme and other schemes. From the analysis of previous researches, Traceback time refers to the requirement time for reconstructing a complete path from Sink to malicious nodes. It is meaningless to traceback for "good" nodes, so this paper just discusses traceback time of malicious nodes. The average credibility for malicious nodes and the average marking probability are given in Fig. 11. The experimental results are consistent with our previous analysis: in the experiment, the proportion for malicious nodes is 10%, the result of trust evaluation function obey the distribution of logarithmic normal distribution: most nodes have high credibility, and a few of nodes trust is low, such as in Fig. 1. The marking probability of nodes with different credibility can be obtained according to Eq. (8) (see from Fig. 3). The marking probability of most trusted nodes are very low, a few of malicious nodes' marking probability is very high (see from Fig. 9). The average marking probability in TAPM scheme is about 19%-27%. For PM scheme, the selected marking probability is the marking probability of the whole network, and in general, most of the selected marking probability is above 40%, but the biggest problem is that it is not targeted. Thus, though the marking probability is about 40%, the marking probability of malicious nodes is only 40%, but the marking probability for malicious nodes in TAPM scheme is about 0.8-1 (see from Fig. 9), it is much higher than that of PM scheme.

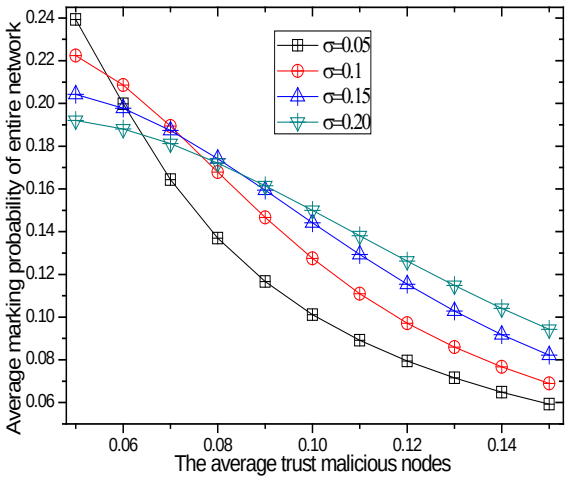


Fig. 11 the average marking probability in TAPM

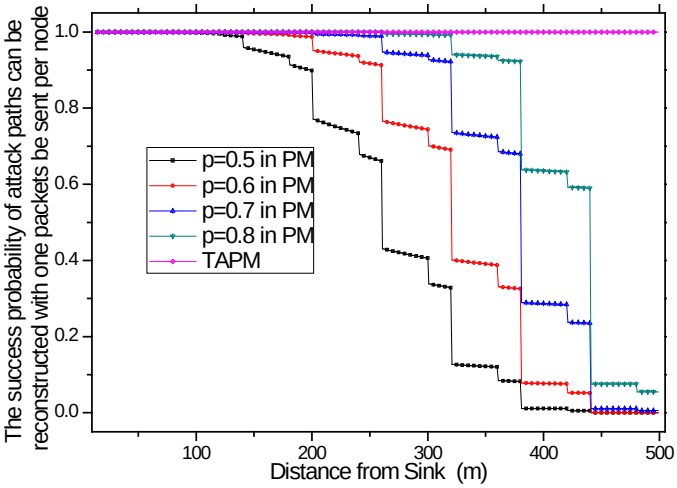


Fig. 12 The success probability of reconstructing attack paths in a round

Fig. 12 shows the success probability of reconstructing attack paths in a round in different distance from the Sink. Experimental parameters in Fig. 12 are: each node sends a data packet to the Sink in the network, TAPM scheme calculates the marking probability used Eq. (8) based on node trust. For PM scheme, the marking probability

is given in Fig. 12. After a round, the probability for reconstructing a complete path from Sink to different nodes is success rate of traceback. In TPAM scheme, the marking probability for malicious nodes is 1, so as long as malicious nodes send a packet to the Sink, the marking information of all nodes in this routing path can be obtained by the Sink, the Sink can reconstruct a complete path when receive a data packet, it may up to 100% traceback. For PM scheme, nodes have the same marking probability in the network, the number of forwarding packets in the area near to the Sink are far more than the number of forwarding packets in the area far from the Sink, so leading to decrease the probability of traceback for nodes far from the Sink (see from Fig. 12).

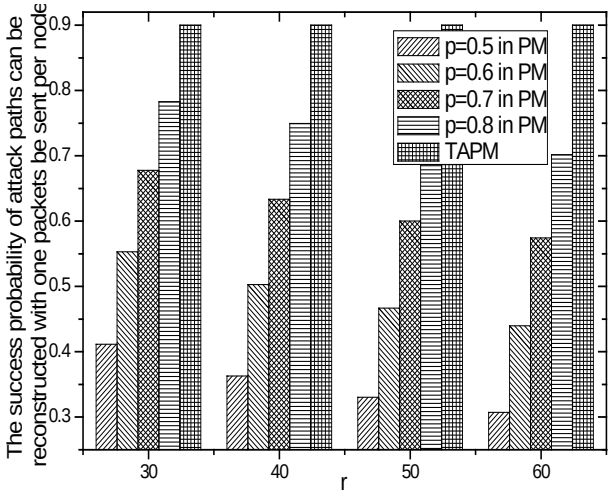


Fig. 13 The average success probability of attack paths can be reconstructed in entire network in a round

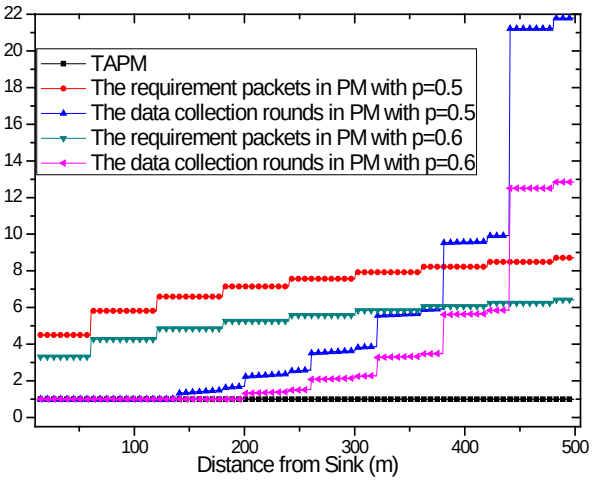


Fig. 14 The traceback time (the data collection rounds need to reconstruct the attack path) with success probability is 0.9

The average probability of success traceback for malicious nodes is given in Fig. 13. In TAPM scheme, it does not guarantee that the marking probability for malicious nodes is high. A small part of malicious nodes have low marking probability, so the average success probability for traceback in the network is about 90%. For PM scheme, because the success probability for traceback in the area far from the Sink is low, and the area of the area is very large, thus the success rate of traceback in the entire network is low. The success rate of traceback is only 33% when the marking probability is 50%, and the success rate of traceback is only 74% when the marking probability is 80%, Fig. 14 shows the requirement traceback time when the traceback success probability is 0.9. The requirement packets refer to the number of marked data packets by node when the traceback success probability is 0.9. The requirement traceback time refer to the requirement rounds when the probability for traceback successful of node is 0.9. It can be seen that traceback time in PM scheme is far more than that of TAPM scheme.

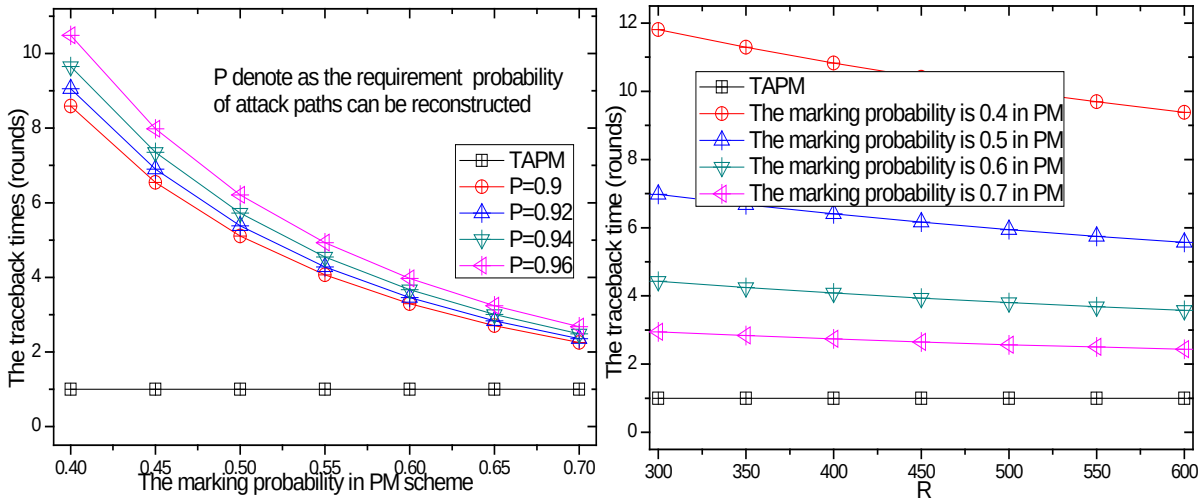


Fig. 15 The traceback time under deference marking probability

Fig. 16 The traceback time (the data collection rounds need to reconstruct the attack path) with success probability is 0.9

Fig. 15, Fig. 16, Fig. 17 give the average traceback time under different marking probability, network radius R and the requirement of success traceback probability of the whole network. It can be seen that the traceback time in TAPM scheme is 1/2-1/10 than that of PM scheme. This shows that TAPM have better traceback ability to protect the network security.

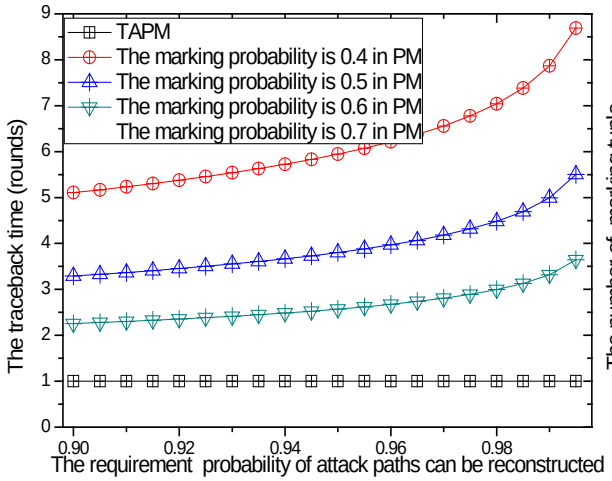


Fig. 17 The average traceback time of attack paths can be reconstructed

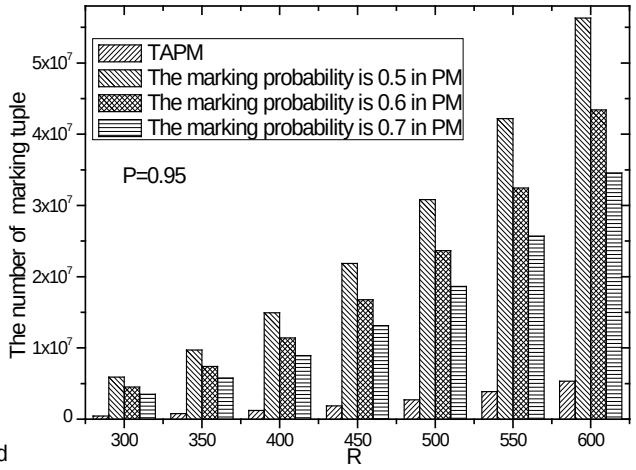


Fig. 18 The number of marking tuple for reconstructing the attack path with success probability is 0.95

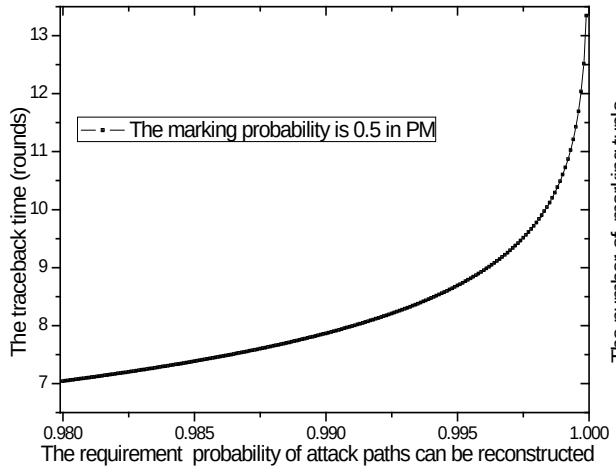


Fig. 19 The traceback time of attack paths can be reconstructed with high requirement probability of attack paths can be reconstructed

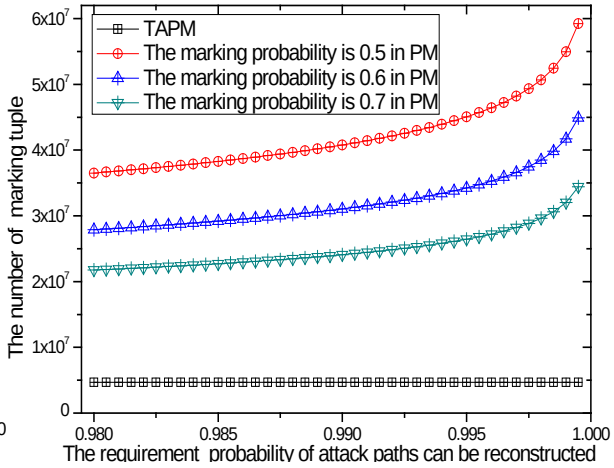


Fig. 20 The number of marking tuple for reconstructing the attack path with high probability

To achieve the specified traceback success probability, The number of produced marking tuples in PM scheme is 6.4-7.5 times than that of TAPM scheme when the marking probability is 0.5 (see from Fig. 18), TAPM scheme has the characteristics of low cost for the system. Especially the required traceback time for PM scheme is rising very fast in the condition of high success probability of traceback (see from Fig. 19), and the system overhead is increasing rapidly (see from Fig. 20), and the cost in TAPM scheme did not change, it shows that TAPM scheme is suitable for security sensitive applications.

6. 2 Analysis of the amount of mark and network lifetime

The number of forwarded data packets and marking tuples in different area under different traceback schemes are given in Fig. 21 and Fig. 22. The results from the experimental show that: (1) the amount of loaded data and marking tuples by nodes in the area near to the Sink is much higher than the amount of loaded data in the area far from the Sink. The change of the number of loaded marking tuples is more intense than the change of the number of data packets, the reason is that: the length of the loaded marking tuples will continue to grow with routing to the Sink, and the length of data packets don't change. (2) In TAPM scheme, the number of loaded marking tuples are far less than the number of loaded marking tuples in PM scheme. This is because a small number of malicious nodes have high marking probability, and most of "good" nodes have low marking probability in TAPM scheme, the average marking probability is still low. (3) The number of loaded data are equal in different schemes, so the difference network lifetime in different scheme are caused by the different of the number of produced marking tuples. So in order to improve the network lifetime, it is important to reduce the number of generated marking tuples by traceback scheme. Because the number of marking tuples in TAPM scheme are far less than that of PM scheme, the network lifetime is longer in TAPM scheme.

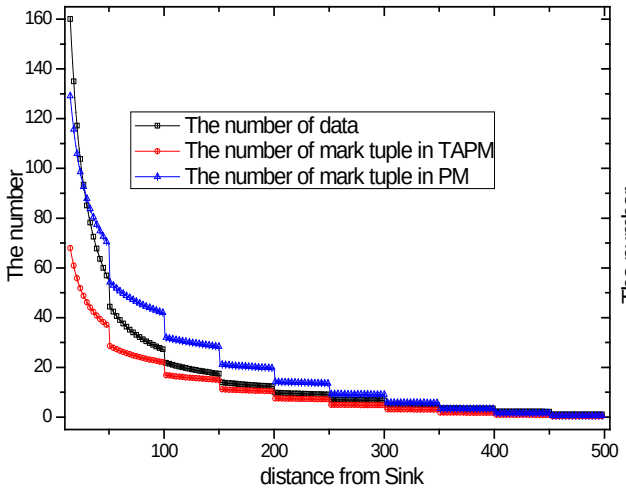


Fig. 21 The marking tuples and under different schemes

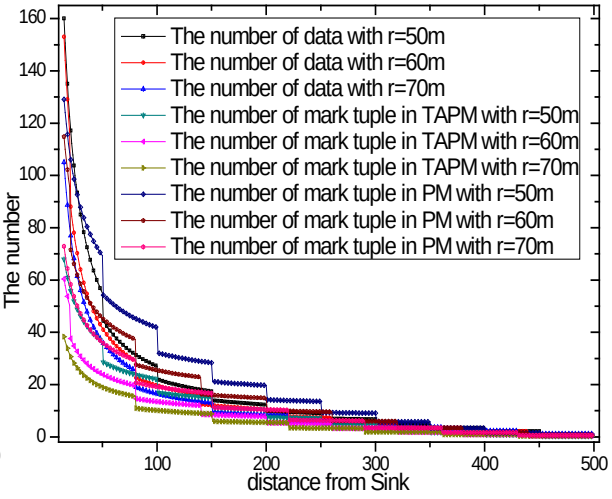


Fig. 22 The marking tuples under different r

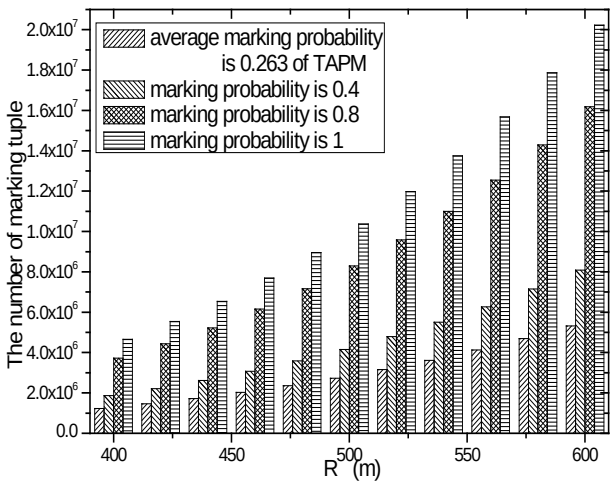


Fig. 23 The total marking tuples under different R

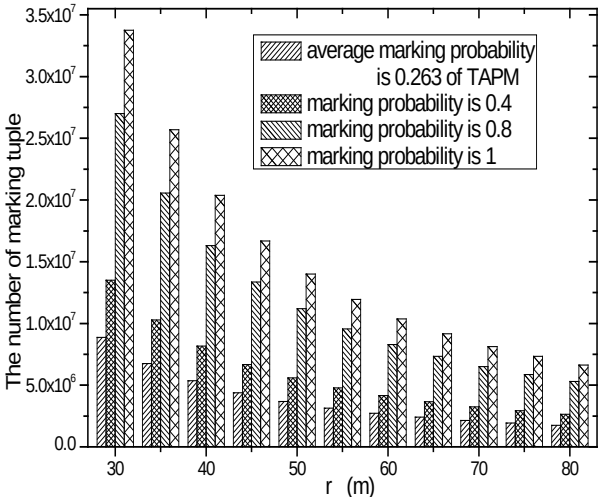


Fig. 24 The total marking tuples under different r

The total marking tuples under different R and r are given in Fig. 23 and Fig. 24 respectively. It can be seen that: (1) the number of total marking tuples in TAPM scheme are far less than that of PM scheme, its performance is

poorer when PM scheme adopts high marking probability, and the performance for TAPM scheme is stable. When r=60m, under different network radius R, the number of total marking tuples are respectively reduced by 34.25%, 67.16% 73.70% than those of PM scheme respectively with marking probability 0.4, 0.8, 1 (see from Fig. 23). The similar conclusions can also be get under different r in 2 schemes. (2) The bigger the network radius R is, the more the number of produced marking tuples are. The bigger the transmitting radius r is, the smaller the number of produced marking tuples are, the reason is: the bigger the r is, the smaller amount of data loaded by each node, which resulting in decreasing the number of loaded marking tuples by each node, at the same time, the number of total marking tuples are decreased. The most extreme is when $r = R$, the number of marking tuples is $\leq$1.

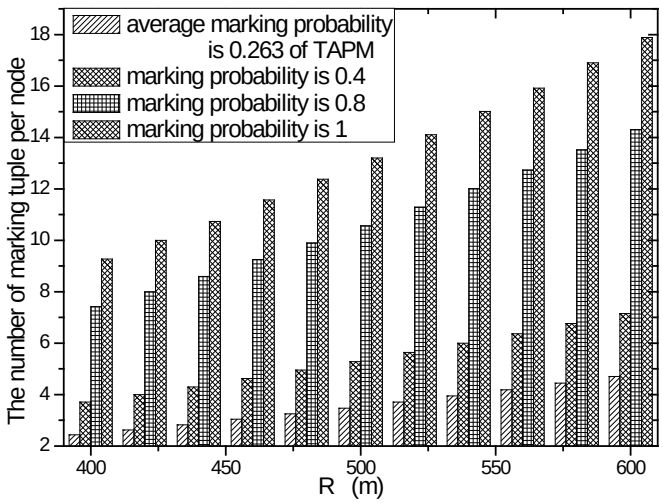


Fig. 25 The marking tuples per node under different R

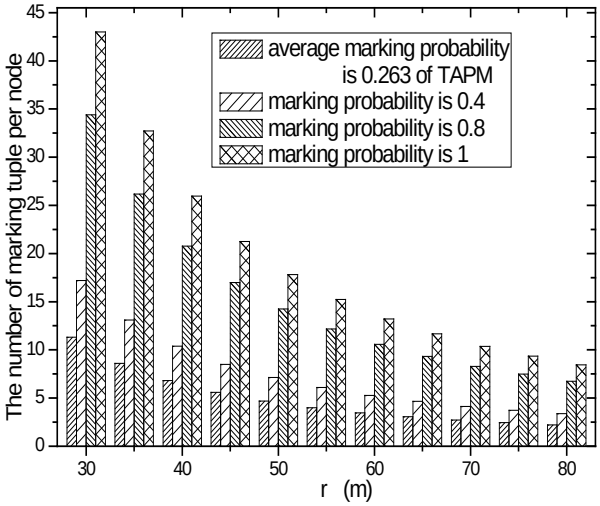


Fig. 26 The marking tuples per node under different r

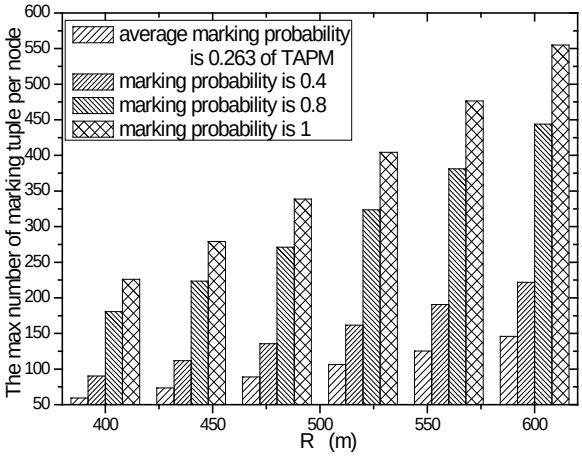


Fig. 27 The max marking tuples per node under different R

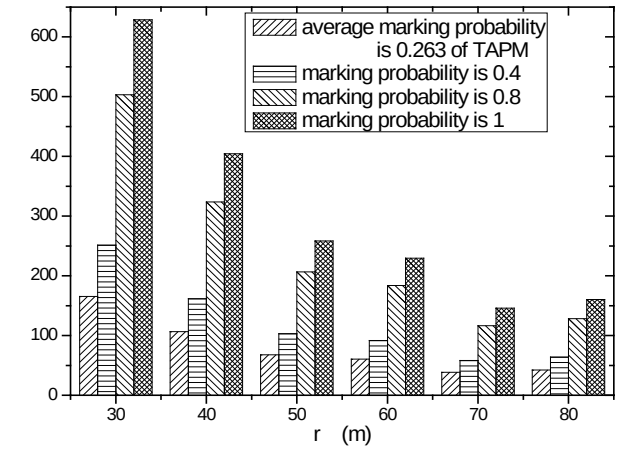


Fig. 28 The max marking tuples per node under different r

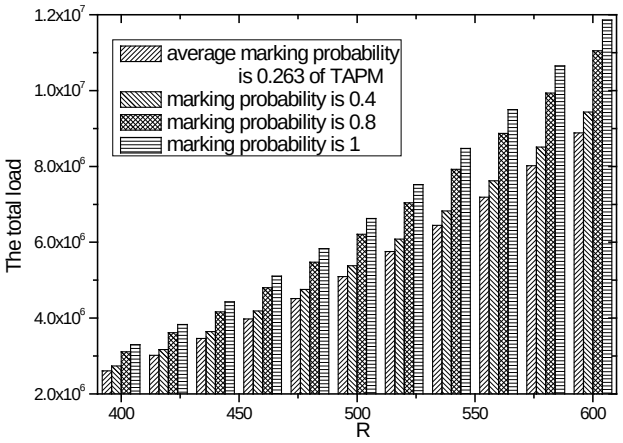


Fig. 29 The sum marking tuples and data under different R with the length of marking tuple is 0.2 times than the length of data packet

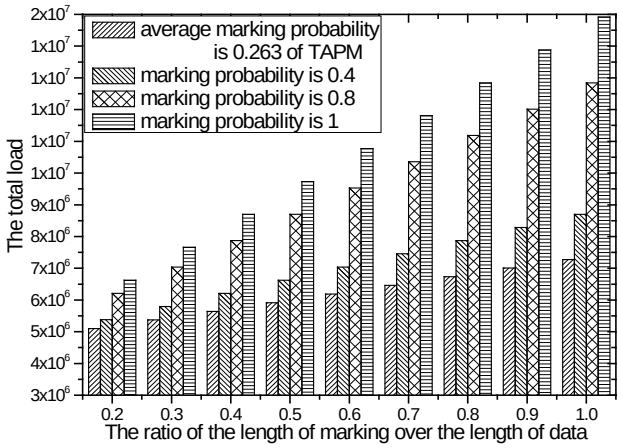


Fig. 30 The sum marking tuples and data under different r with the length of marking tuple is 0.2 of the length of data packet

The number of assumed average marking tuples by each node under different R and r are given in Fig. 25 and Fig. 26, respectively. Due to the total number of marking tuples are determined in Fig. 25 and Fig. 26, the number of loaded average marking tuples is refer to its total number of marking tuples divided by the number of nodes in the network, so the results in Fig. 25 and Fig. 26 is similar to the results in Fig. 23 and Fig. 24. Because the network lifetime is the time of the death of the first node in the network, and the number of loaded data packets are equal in different traceback schemes, so the different network lifetime under different schemes are caused by the largest number of marking tuples. Fig. 27 and Fig. 28 give the maximum number of loaded marking tuples by nodes, which can be seen from the experimental, the maximum number of marking tuples in TAPM scheme are far less than that of PM scheme.

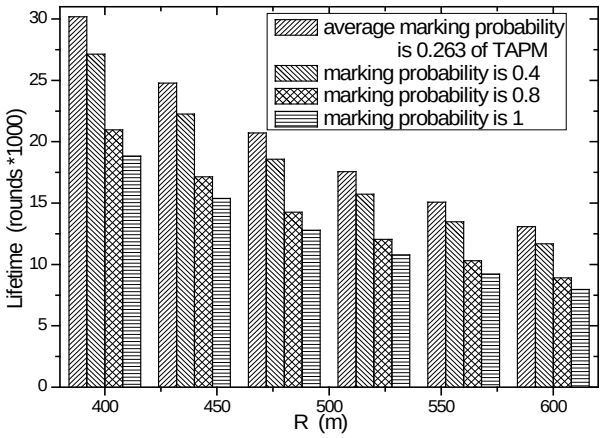


Fig. 31 The lifetime under different R with the length of marking tuple is 0.6 of the length of data packet

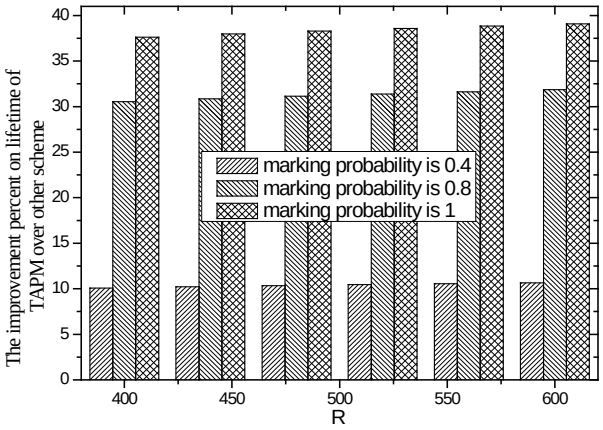


Fig. 32 The improvement percent on lifetime of TAPM over other scheme under different R

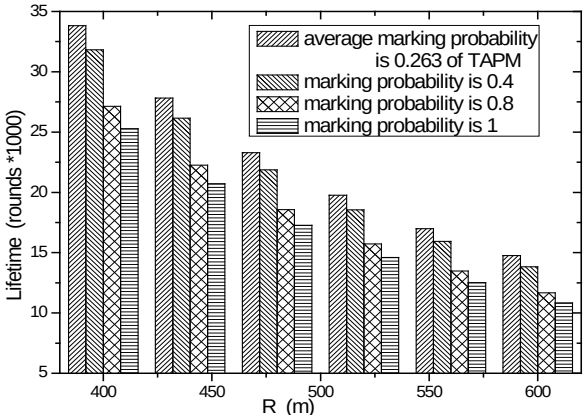


Fig. 33 The lifetime under different R with the length of marking tuple is 0.3 of the length of data packet

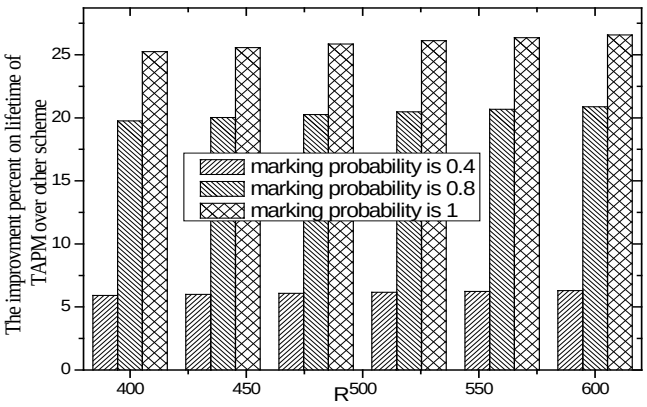
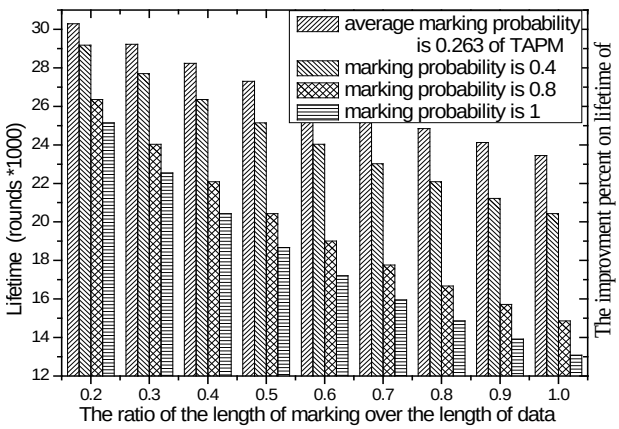


Fig. 34 The improvement percent on lifetime of TAPM over other scheme under different R

Fig. 29 and Fig. 30 given the total amount of marking information with the length of marking tuple is 0.2 times than the length of data packet in different R and r. The total information = the number of data packets + 0.2* the number of marking tuples. It can be seen from the experimental results that the number of data packet are equal in different schemes. The difference of the number of marking tuple in different scheme is very large, but due to the length of marking tuple is 0.2 times than the length of data packet, the difference in different schemes is became small from the results of Fig. 29 and Fig. 30. It can be seen from Fig. 31 and Fig. 32 that: the improvement percent on lifetime of TAPM over PM scheme is 10%, 30%, 37% with the length of marking tuple is 0.6 of the length of data packet. However, when the length of marking tuple is 0.3 of the length of data packet, the improvement percent on lifetime of TAPM over PM scheme is 6%, 19%, 25% (see from Fig. 33 and Fig. 34). The network lifetime can be seen from Fig. 35 with difference ratio of the length of marking tuple over the length of data packet (denote as $\emptyset$). We can see from Fig. 35 that: the more the $\emptyset$ is, the smaller the network lifetime is. This is because, the greater the $\emptyset$ is, the more the amount of information forwarded by nodes, which lead to the lower network lifetime. However, the larger the $\emptyset$ is, the bigger the proportion for marking tuple account for the forwarded information is. Due to the less number of marking tuple in TAPM scheme, so the advantages of the network lifetime are more obvious than other schemes. From Fig. 36, you can see that as the $\emptyset$ increases from 0.2 to 1, the network lifetime is improved by 3.66%-12.84 in TAPM scheme than that of PM scheme with 0.4 marking probability, and the network lifetime is improved by 12.99%-36.61% in TAPM scheme than that of PM scheme with 0.8 marking probability, the network lifetime is improved by 17.05%-44.26% in TAPM scheme than that of PM scheme with 1 marking probability. Compare to PM scheme, TAPM scheme has greatly improve on security (traceback time) and the network lifetime. In general, to improve the safety of traceback scheme, it often needs to consume more energy. But the two key performance indicators are simultaneously improved in TAPM scheme in this paper, this is a kind of innovation.



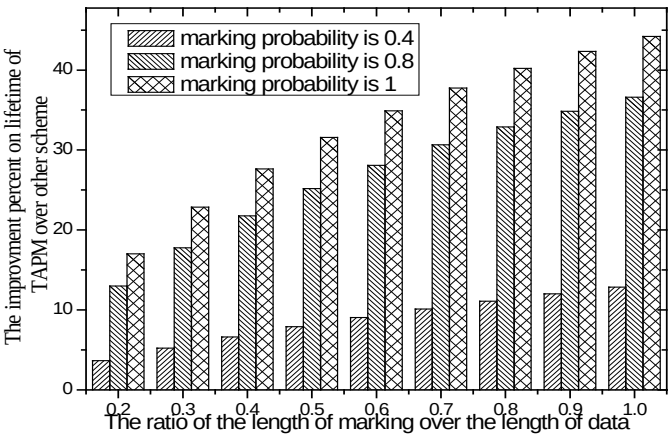


Fig. 35 The lifetime

Fig. 36 The improvement percent on lifetime of TAPM over other scheme

6. 3The performance affected by parameters

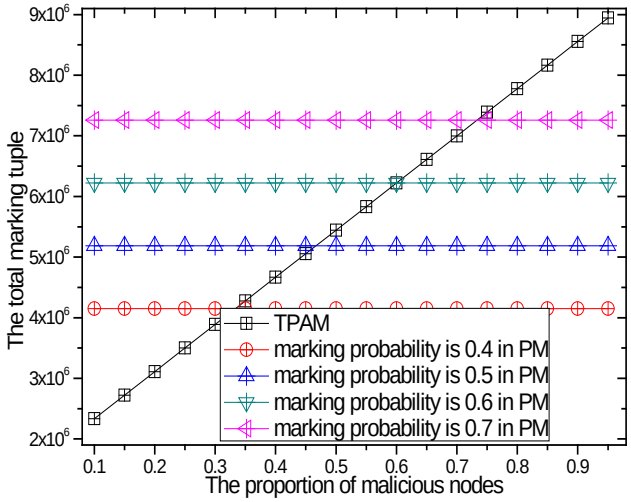


Fig. 37 The amount of marks affected by the proportion of malicious nodes

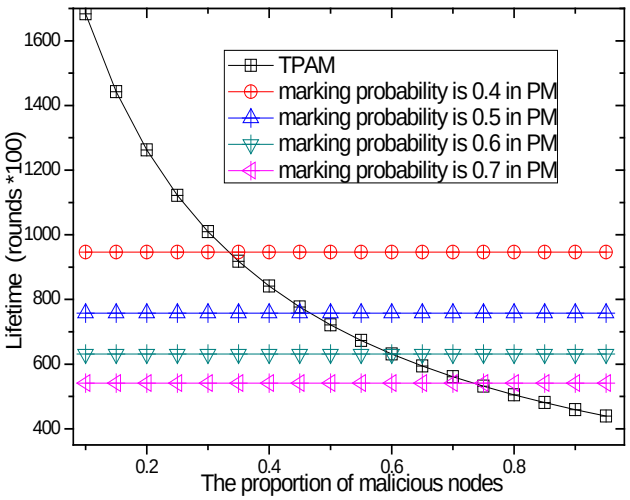


Fig. 38 The lifetime affected by the proportion of malicious nodes

An important premise in TAPM scheme is: the proportion of malicious nodes is small in the network, such as less than 15%, it meets the practical situation. The number of marking tuples and network lifetime are given under different proportion of malicious nodes in Fig. 37and Fig. 38. It can be seen that with the increase of the proportion of malicious nodes, the number of marking tuples are straight up, while the network lifetime is decreased in TAPM scheme. It explain that TAPM scheme can only be used to the less proportion of malicious nodes in the networks. And PM scheme have nothing to do with malicious nodes.

The effect of the variance of malicious nodes' credibility on network performance are given in Fig. 39 and Fig. 40. When the average credibility of nodes is 0.05 (low), because the marking probability for nodes with low credibility is high, if the variance of malicious nodes' credibility increase, the range in 0.05 of the distribution of malicious nodes' credibility are increased. However, the marking probability is decreased when nodes' credibility is large, the space for reducing marking probability is only 0.05, that is to say the growth space is limited. Fig. 3 shows that the network lifetime increases with the decrease of the average marking probability (see from Fig. 40). The 0.1 variance and 0.15 variance of malicious nodes' credibility is just the opposite. The experiment shows that the ratio of malicious nodes and density distribution have effect on the performance of TAPM scheme.

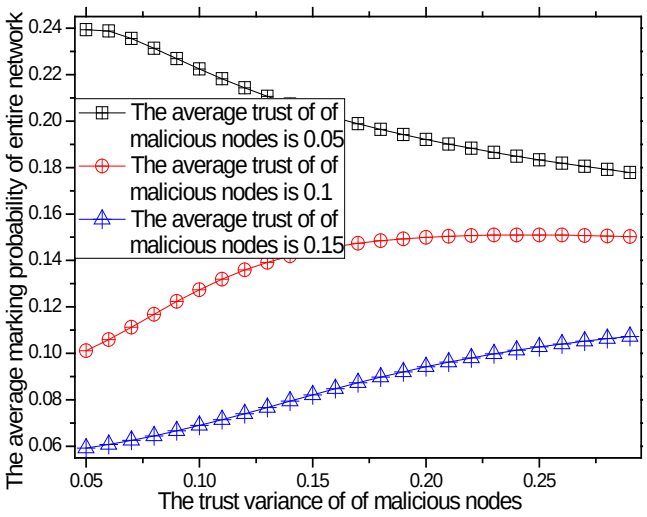


Fig. 39 The average marking probability affected by the trust variance

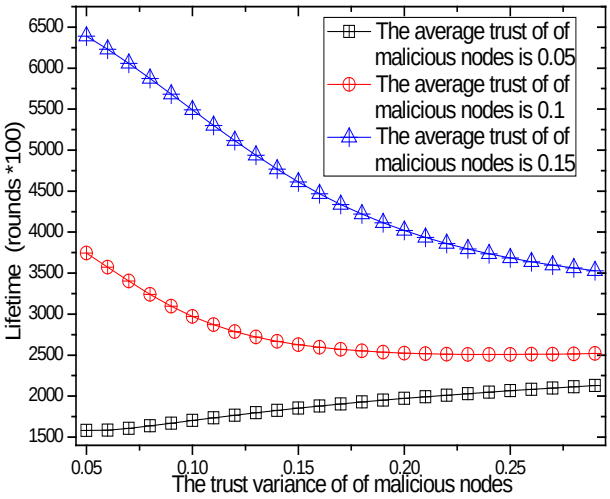


Fig. 40 The lifetime affected by the trust variance

7 conclusion

In this paper, a trust-aware probability marking (TAPM) traceback scheme is proposed to fast locate malicious source and guarantee cyber security. TAPM scheme sets different marking probability based on node trust which is difference from the same marking probability of nodes in the previous schemes. In the network, it is sensible to rebuilt attach path of malicious nodes, so it adopts high marking probability for nodes with low credibility to reduce the traceback time, which is advantageous to network security. A small part of network nodes are malicious nodes, though the malicious nodes have high marking probability, the average marking probability of the entire network is still low, TAPM scheme can reduce the amount of marking information, so as to improve the network lifetime. TAPM scheme not only improve the network lifetime, but also can reduce the traceback time. Moreover, TAPM scheme adopts active detection methods to ensure that the attach path can be reconstructed. The results of experiments have proved that TAPM scheme has good performances. TAPM respectively reduce the number of the total marking tuples by 34.25%, 67.16%, 73.70% than PM scheme respectively with the 0.4, 0.8, 1 marking probability, the network lifetime respectively is improved by 3.66%-12.84,12.99%-36.61%,17.05%-44.26% than PM scheme, the trace time is 1/2-1/10 times than that of PM scheme. Though TAPM scheme has better performance, the energy utilization is lower. The areas far away from sink have much more energy left, but the energy consumption in the areas near to the Sink is highest. In the future work, a scheme, makes full use of the left energy and has better performance, should be studied.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61379110, 61073104, 61572528, 61272494, 61572528), The National Basic Research Program of China (9703 Program) (2014CB046305), JSPS KAKENHI Grant Number 25880002, and JSPS A3 Foresight Program, Fundamental Research Funds for the Central Universities of Central South University (2015zzts215).

References

[1] J. Jang-Jaccard, S. Nepal, A survey of emerging threats in cybersecurity, Journal of Computer and System Sciences. 80(5) (2014) 973-993.

[2] W. Bul'ajoul, A. James, M. Pannu, Improving network intrusion detection system performance through quality of service configuration and parallel technology, Journal of Computer and System Sciences. 81(6) (2015) 981-999.

[3] BR. Ray, J. Abawajy, M. Chowdhury, Scalable RFID security framework and protocol supporting Internet of Things, Computer Networks 67 (2014) 89-103.

[4] IRA. Hamid, JH. Abawajy, An approach for profiling phishing activities, Computers & Security 45 (2014) 27-41.

[5] B. Borgman, S. Mubarak, KKR. Choo, Cyber security readiness in the South Australian Government, Computer Standards & Interfaces 37 (2015) 1-8.

[6] NAM. Junghyun, KKR. CHoo, P. Juryon, W. Dongho, An Offline Dictionary Attack against Abdalla and Pointcheval's Key Exchange in the Password-Only Three-Party Setting, IEICE Transactions Fundamentals of Electronics, Communications and Computer Sciences, 98(1) (2015) 424-427.

[7] Y. Hu, A. Liu, Improvement the quality of mobile target detection through portion of node with fully duty cycle in WSNs. Computer Systems Science and Engineering, 31 (1) (2016) 5-17, 2016.

[8] Y. Liu, M. Dong, K. Ota, A. Liu. ActiveTrust: Secure and Trustable Routing in Wireless Sensor Networks. IEEE Transactions on Information Forensics and Security, 11(9) (2016) 2013-2027. DOI: 10.1109/TIFS.2016.2570740, 2016.

[9] S. M. Alam, S. Fahmy, A practical approach for provenance transmission in wireless sensor networks, Ad Hoc Networks. 16 (2014) 28-45.

[10] B. C. Cheng, H. Chen, Y. J. Li, et al, A packet marking with fair probability distribution function for minimizing the convergence time in wireless sensor networks, Computer Communications. 31(18) (2008) 4352-4359.

[11] M. S. Siddiqui, S. Obaid Amin, C. S. Hong, Hop-by-hop traceback in wireless sensor networks, IEEE communications letters. 16(2) (2012) 242-245.

[12] J. Xu, X. Zhou, F. Yang, Traceback in wireless sensor networks with packet marking and logging, Frontiers of Computer Science in China. 5(3) (2011) 308-315.

[13] Y. Hu, X. Mian, K. Ota, A. Liu, M. Guo, Mobile Target Detection in Wireless Sensor Networks with Adjustable Sensing Frequency, IEEE System Journal, DOI: 10.1109/JSYST.2014.2308391, 2014.

[14] T. Y. Wong, K. T. Law, J. C. S. Lui, et al, An efficient distributed algorithm to identify and traceback DDos traffic, The Computer Journal. 49(4) (2006) 418-442.

[15] O. Osanaiye, KKR. Choo, M. Dlodlo, Distributed denial of service (DDoS) resilience in cloud: Review and conceptual cloud DDoS mitigation framework, Journal of Network and Computer Applications, DOI: 10.1016/j.jnca.2016.01.001, 2016.

[16] J. Nam, KKR. Choo, S. Han, M. Kim, J. Paik, D. Won, Efficient and Anonymous Two-Factor User Authentication in Wireless Sensor Networks: Achieving User Anonymity with Lightweight Sensor Computation. PLoS ONE. DOI:10.1371/journal.pone.011670910, 2015.

[17] E. H. Jeong, B. K. Lee, An IP Traceback Protocol using a Compressed Hash Table, a Sinkhole router and data mining based on network forensics against network attacks, Future Generation Computer Systems. 33 (2014) 42-52.

[18] S. Saurabh, A. S. Sairam, ICMP based IP traceback with negligible overhead for highly distributed reflector attack using bloom filters, Computer Communications. 42 (2014) 60-69.

[19] N. Lu, Y. Wang, S. Su, et al, A novel path-based approach for single-packet IP traceback, Security and Communication Networks. 7(2) (2014) 309-321.

[20] M. Ge, KKR. Choo, A Novel Hybrid Key Revocation Scheme for Wireless Sensor Networks. In: Proceedings of 8th International Conference on Network and System Security, 2014, pp. 462-475.

[21] M. Ge, KKR. Choo, H. Wu, Y. Yu, Survey on key revocation mechanisms in wireless sensor networks, Journal of Network and Computer Applications, 63 (2016) 24-38.

[22] Y. Tseng, H. Chen, W. Hsieh, Probabilistic packet marking with non-preemptive compensation, IEEE Communications Letters. 8(6) (2004) 359-361.

[23] Yang L, Cao J, Zhu W, et al, Accurate and Efficient Object Tracking based on Passive RFID, IEEE Transactions on Mobile Computing. 14(11) (2015) 2188-2200.

[24] X. Liu, K. Ota, A. Liu, Z. Chen, An incentive game based evolutionary model for crowd sensing networks, Peer-to-Peer Networking and Applications, 9(4) (2016) 692-711. DOI: 10.1007/s12083-015-0342-2.

[25] X. Liu, M Dong, K. Ota, P. Hung, A. Liu. Service Pricing Decision in Cyber-Physical Systems: Insights from Game Theory, IEEE Transactions on Services Computing, 9 (2) (2016) 186-198. DOI: 10.1109/TSC.2015.2449314.

[26] Dai H, Chen G, Wang C, et al. Quality of energy provisioning for wireless power transfer, IEEE Transactions on Parallel and Distributed Systems. 26(2) (2015) 527-537.

[27] M. Dong, K. Ota, L. T. Yang, et al, LSCD: A Low Storage Clone Detecting Protocol for Cyber-Physical Systems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 35(5) (2016) 712-723. DOI: 10.1109/TCAD.2016.2539327.

[28] A. Aburumman, W.j. Seo, R. Islam, M.k. Khan, KKR. Choo, A Secure Cross-Domain SIP Solution for Mobile Ad Hoc Network Using Dynamic Clustering, In: Proceedings of 11th International Conference on Security and Privacy in Communication Networks, 2015, pp. 649-664.

[29] A. Aburumman, KKR. Choo, A Domain-Based Multi-cluster SIP Solution for Mobile Ad Hoc Network, In: Proceedings of 10th International ICST Conference on Security and Privacy in Communication Networks, 2015, pp. 267-281.

[30] D. R. Waldo, L. W. Smith, E. L. Cox, et al, Logarithmic normal distribution for description of sieved forage materials, Journal of Dairy Science. 54(10) (1971) 1465-1469.

[31] G. Han, J. Jiang, L. Shu, J. Niu, HC. Chao. Management and applications of trust in Wireless Sensor Networks: A survey, Journal of Computer and System Sciences. 80(3) (2014) 602-617.

[32] A. Liu, X. Liu, Y. Liu. A comprehensive analysis for fair probability marking based traceback approach in WSNs. Security and Communication Networks. DOI: 10.1002/sec.1515, 2016.