Identifying the propagation sources of stealth worms*

Yanwei Sun[1,2,3], Lihua Yin[1], Zhen Wang[4](✉),
Yunchuan Guo[2,3], and Binxing Fang[5]

[1] Cyberspace Institute of Advanced Technology (CIAT), Guangzhou University, Guangzhou, China

[2] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[3] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

[4] School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China
wangzhen@hdu.edu.cn

[5] Institute of Electronic and Information Engineering of UESTC in Guangdong, China

Abstract. Worm virus can spread in various ways with great destructive power, which poses a great threat to network security. One example is the WannaCry worm in May 2017. By identifying the sources of worms, we can better understand the causation of risks, and then implement better security measures. However, the current available detection system may not be able to fully detect the existing threats when the worms with the stealth characteristics do not show any abnormal behaviors. This paper makes two key contributions toward the challenging problem of identifying the propagation sources: 1) A modified algorithm of observed results based on Bayes rule has been proposed, which can modify the results of possible missed nodes, so as to improve the accuracy of identifying the propagation sources. 2) We have applied the method of branch and bound, effectively reduced the traversal space and improved the efficiency of the algorithm by calculating the upper and lower bounds of the infection probability of nodes. Through the experiment simulation in the real network, we verified the accuracy and high efficiency of the algorithm for tracing the sources of worms.

Keywords: stealth worm, propagation sources, Bayes rule

1 Introduction

Worms are spreading rapidly via emails, social networks and self-scanning etc. Moreover, they are also very destructive. In May 2017, the WannaCry worm

* This work was supported by the National Key R&D Program of China (No.2016YFB0800702), Natural Science Foundation of Zhejiang Province (Grant Nos. LY18F030007 and LY18F020017), and DongGuan Innovative Research Team Program (No.201636000100038)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_79

broke out worldwide via MS17-010 bug, infected at least 200,000 users in 150 countries[1], and resulted in the losses of almost 4 billion USD[2]. In order to effectively prevent the spreading of worms, the most critical means is to identify the source of spreading [1]. However, there is high false negative rate of the monitoring results because some worms are exploiting zero-day vulnerabilities and some are changing their own characteristics to avoid the detection [2]. So it is a great challenge to identify the sources of stealth worms in the case of high false negative rate.

To infer the origin of the propagation, one of the principles is to employing the maximum likelihood estimation on each potential source, and then select the most likely one as the propagation source. But these studies ignored the stealth characteristic, which may cause a false negative rate for the observing results, for instance, the false negative rate of honeycyber [3] reached about 0.92%. In this case, the results acquired by the existing identifying methods may have deviation.

Aiming at this problem, this paper mainly discusses the methods for identifying the sources of stealth worms. We use Bayesian Theory to correct the observed results of each node, and then propose an efficient algorithm based on branch and bound. Experimental results on three real-world data sets empirically demonstrate that our method consistently achieves an improvement in accuracy.

2 Related work

There are many representative studies on the issue of identifying the propagation source. According to the differences in the observations, we can divide the study into complete observation [4] [5] and snapshot [6] [7] [8]. In order to find the rumor source, Shah et al. [4] constructed a maximum likelihood estimator based on SIR model, and then proposed a computationally efficient algorithm to calculate the rumor centrality for each node. Fioriti et al. [5] focused on locating the multiple origins of a disease outbreak. When a node had been removed, the larger the reduction of the eigenvalue, the more likely this node was the origin. Compared with complete observation, snapshot provided less information and attracted a lot research. Prakash et al. [6] proposed a two step approach to identify the number of seed nodes based on SI model. They first found the high-quality seeds and then calculated the minimum description length score to identify the best set of seeds. Lokhov et al. [7] defined the snapshot as the following case: there was only single source at initial time t and the observation was conducted at t_0 where $t_0 - t$ was unknown. By discussing the propagation dynamic equations, the DMP method chose the node which had the highest

[1] http://www.straitstimes.com/tech/accidental-hero-halts-global-ransomware-attack-but-warns-this-is-not-over

[2] https://www.cbsnews.com/news/wannacry-ransomware-attacks-wannacry-virus-losses

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_79

probabilities that could produce the snapshot. Luo et al. [8] dealt with the single source at SIS model, and showed that the source estimator was a Jordan infection center. However, all these works ignored the stealth characteristic. In this paper, we mainly discuss the methods for identifying the sources of stealth worms.

3 Identify the propagation resource

3.1 Basic assumptions

Worm propagation in our study follows SI model, also we use discrete time model and assume that it takes one time tick to infect a suspectable node. After the end of each time tick, we record the monitor result and record the infected time if the node is infected. We use the directed graph $G = (V, E)$ to represent the network topology in which $V = \{1, 2, \ldots, n\}$ is the set of nodes and E is the set of edges. More specifically, we use V_S for suspectable node set and V_I for infected node set. If $(i, j) \in E$ and $i \in V_I$, $j \in V_S$, we use r_{ij} to denote the probability that node j is infected by node i. We assume r_{ij} is a fixed value, the quantization process is beyond the scope of this article, we assume that this value is known.

The network is observed over a time period $[0, T]$. For the uninfected nodes in the observation, the real status may be uninfected, or may be infected but undetected. For the infected nodes in the observation, we assume that the real state is infected. However, the infection time recorded in the observation result only indicates that the node is found infected at that time, which may not be the actual infection time of the node. Altogether, we consider the situation that the detection technique may has false negative rate and has no false positive rate.

3.2 Identify the propagation source

The process of correction. Let's illustrate the correction process through an example. The network shown in fig.1 has total 9 nodes. At time tick t, the node 2,3,6,7 are detected infected. According to the network connectivity, we conclude that at this time, node 8 (labeled yellow) has a high probability of being a false negative (the probability calculation will be introduced at next subsection). Since a new node is considered to be a infected node and has the ability to infect other nodes, we have to re-traverse the remaining uninfected nodes. In the next traversal process, because of the influence of adding node 8, it is estimated that the probability of node 5 being infected also exceeds the threshold, so node 5 is considered as the infected node. The above process is repeated until no new node is found infected and then the traversal at time t ends.

At time $t + 1$, observations show that node 4 is found to be infected, so we traverse node 1 and node 9, finding that the probabilities of false negative of both two nodes are pretty low, so we believe node 1 and node 9 are not infected at time $t + 1$. The traversal at time $t + 1$ ends.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_79

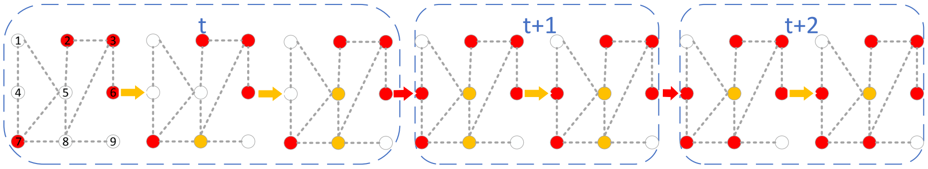


Fig. 1. An illustrative example

At time $t + 2$, it is observed that node 8 is infected. Since node 8 has previously been identified as an infected node, this shows that there is a delay in the observation results in terms of node 8. At this point, the observation results tend to be stable, i.e. nodes 1, 9 are uninfected nodes, node 5 is infected but undetected nodes, and the remaining nodes are infected and detected nodes.

Calculate the false negative probability. We use j_{obv}^t to represent the observation at time tick t for node j. For each $j_{obv}^t \in V_S^t$, we fist compute the probability of being infected:

$$P(j \in V_I^t) = 1 - \prod_{(i,j)\in E \wedge i \in V_I^t} (1 - r_{ij}) \tag{1}$$

After obtaining the above probability, we calculate the probability that node j is in an infected state under the condition that its observation result is uninfected by using the Bayesian formula:

$$\begin{aligned} P(j \in V_I^t | j_{obv}^t \in V_S^t) &= \frac{P(j \in V_I^t \wedge j_{obv}^t \in V_S^t)}{P(j_{obv}^t \in V_S^t)} \\ &= \frac{P(j \in V_I^t) \cdot P(j_{obv}^t \in V_S^t | j \in V_I^t)}{P(j \in V_I^t) \cdot P(j_{obv}^t \in V_S^t | j \in V_I^t) + P(j \in V_S^t) \cdot P(j_{obv}^t \in V_S^t | j \in V_S^t)} \\ &= \frac{P(j \in V_I^t) \cdot P_{FN}}{P(j \in V_I^t) \cdot P_{FN} + (1 - P(j \in V_I^t)) \cdot (1 - P_{FN})} \end{aligned} \tag{2}$$

We assume that P_{FN} is a fixed value and is known. This assumption is reasonable, because this value can be obtained from the statistics of past observations and real results. So we first compute $P(j \in V_I^t)$ and then $P(j \in V_I^t | j_{obv}^t \in V_S^t)$. If the above probability exceeds our preset threshold Th, for example 80%, then we think that the observation of the state of node j is wrong.

After correcting the observation, we use DMP algorism[7] to infer the origin of the propagation. As for the algorithm itself, this article will not go into details.

4 An efficient traversal algorithm for correction process.

During the process of correction, every time a missed node is found, it is necessary to re-execute the iteration. If the algorithm is directly applied to large-scale networks, its efficiency is not satisfactory. Therefore, we optimize the iterative process of the algorithm based on branch and bound:

Algorithm 1 Revising the observed result

```
Input:
    Network $G$; the infection probability $r_{ij}$; observed result set; the threshold $Th$.
Output:
    A set of revised result.
Initialize $V_S^0$ and $V_I^0$ based on observed result; Initialize $V_{FN}^t = V_{Remove}^t = \emptyset$
for each time point $t \in [0, T]$ do
    while repeat time $t_r \leq K$ do
        for each node $j \in V_S^t$ do
            Calculating $P(j \in V_I^t)$ and $P(j \in V_I^t | j_{obv}^t \in V_S^t)$
            if $P(j \in V_I^t | j_{obv}^t \in V_S^t) \geq Th$ then
                Set $V_I^t = V_I^t \bigcup j$ and $V_S^t = V_S^t \setminus j$ and $V_{FN}^t = V_{FN}^t \bigcup j$
                Record the infected time of node $j$
            else
                Calculating $P_{max}(j \in V_I^t)$ and $P_{max}(j \in V_I^t | j_{obv}^t \in V_S^t)$
                if $P_{max}(j \in V_I^t | j_{obv}^t \in V_S^t) \leq Th$ then
                    Set $V_S^t = V_S^t \setminus j$ and $V_{Remove}^t = V_{Remove}^t \bigcup j$
                end if
            end if
        end for
        if The number of infected nodes is larger then
            Sort the $V_S^t$ according to the number of infected neighbors DESC.
        else
            Sort the $V_S^t$ according to the number of neighbors ASC.
        end if
    end while
end for
return $V_I^t$ and $V_S^t$ and $V_{FN}^t$ and $V_{Remove}^t$
```

Algorithm 1 shows the efficient traversal method for correction process. The optimization idea is as follows: at time t, for the suspectable nodes in the observation result, if any node satisfies the following condition, the node must be an uninfected node:

$$\begin{aligned} P_{max}(j \in V_I^t | j_{obv}^t \in V_S^t) &= \frac{P_{max}(j \in V_I^t \wedge j_{obv}^t \in V_S^t)}{P_{max}(j_{obv}^t \in V_S^t)} \\ &= \frac{P_{max}(j \in V_I^t) \cdot P_{FN}}{P_{max}(j \in V_I^t) \cdot P_{FN} + (1 - P_{max}(j \in V_I^t)) \cdot (1 - P_{FN})} \leq Th \end{aligned} \tag{3}$$

where:

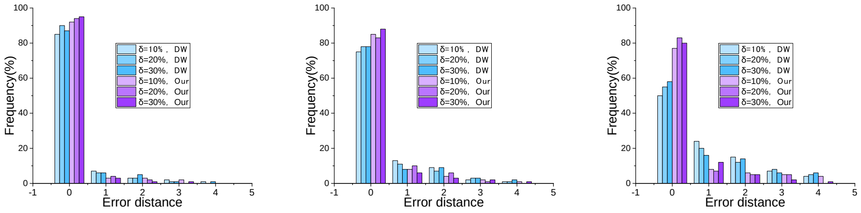
$$P_{max}(j \in V_I^t) = 1 - \prod_{(i,j) \in E} (1 - r_{ij}) \tag{4}$$

It can be seen that in the calculation, the number of neighbors around node j is relaxed. The idea is that even if all its neighbors are infected, the probability of the node j being infected is still small, so that the $P_{max}(j \in V_I^t | j_{obv}^t \in V_S^t)$ is less than the pre-set threshold. It can be concluded that this node is an uninfected node, regardless of its neighbors' real state.

Also, in order to reduce the iteration round as much as possible, we do not terminate the traversal immediately after discovering an missing node at each iteration, but move the node from set V_S^t to V_I^t and then continue traversing subsequent nodes. Inspired by this idea, we need to adjust the traversal order of these two traversals. More specifically, the node which has more infected neighbors has the higher priority to traverse, because it is more likely to be the false negative node compared with other nodes. Meanwhile, we could also traverse the node which has less neighbors, because this node is more likely to be the real uninfected one. Which way to choose is depend on the propagation situation. If the number of infected node is larger than the suspectable node, that means the worm spreads rapidly, and we should choose the first way to traverse.

5 Experement.

The method proposed in this paper was tested on three real world networks, include the power grid network[3], the enron email network[4] and AS-level network[5]. For the sake of discussion, the infection probability between nodes in the above networks were generated randomly, and it was assumed that $r_{ij} = r_{ji}$. All experiments were subject to independent performance test in windows7 system. The test computer was configured as an Intel Core i7-6700 3.4GHz processor, 8 GB memory and 4G virtual memory allocated by ECLIPSE.



(a) Enrom emali network (b) Power grid network (c) AS-level network

Fig. 2. Accuracy comparison with the existing work.

Accuracy comparison with the existing work. The accuracy between this work and [9] was compared. We used the same configurations with that work. The

[3] http://www-personal.umich.edu/ mejn/netdata/

[4] http://www.cs.cmu.edu/ enron/

[5] http://data.caida.org /datasets/as-relationships/serial-1/

false negative rate was not considered in [9], so when the false negative rate was higher, the accuracy of [9] was decreased significantly, and the effect of this work was significantly better in this case. It can be seen that, when the false negative rate was close to 20%, the probability of error distance=0 for [9] was only 50%, while it was remained at about 70% in our algorithm.

6 Conclusion.

This paper presents the first work on identifying the propagation source of stealth worm. We propose a modified algorithm of observed results based on bayes formula, which can modify the results of possible false negative nodes, so as to improve the accuracy of identifying the propagation sources. After that, we have applied the method of branch and bound, effectively reduced the traversal space and improved the efficiency of the algorithm by calculating the upper and lower bounds of the infection probability of nodes. We test our algorithm on three real networks ,and the results show the accuracy of the algorithm.

References

1. Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, 19(1):465–481, 2017.
2. Ratinder Kaur and Maninder Singh. A survey on zero-day polymorphic worm detection techniques. *IEEE Communications Surveys & Tutorials*, 16(3):1520–1549, 2014.
3. Mohssen MZE Mohammed, H Anthony Chan, and Neco Ventura. Honeycyber: Automated signature generation for zero-day polymorphic worms. In *Military Communications Conference, 2008. MILCOM 2008. IEEE*, pages 1–6. IEEE, 2008.
4. Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8):5163–5181, 2011.
5. Vincenzo Fioriti and Marta Chinnici. Predicting the sources of an outbreak with a spectral technique. *arXiv preprint arXiv:1211.2333*, 2012.
6. B Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. Efficiently spotting the starting points of an epidemic in a large graph. *Knowledge and information systems*, 38(1):35–59, 2014.
7. Andrey Y Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E*, 90(1):012801, 2014.
8. Wuqiong Luo and Wee Peng Tay. Finding an infection source under the sis model. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2930–2934. IEEE, 2013.
9. Derek Wang, Sheng Wen, Yang Xiang, Wanlei Zhou, Jun Zhang, and Surya Nepal. Catch me if you can: Detecting compromised users through partial observation on networks. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*, pages 2417–2422. IEEE, 2017.