RAFM: A Real-time Auto Detecting and Fingerprinting Method for IoT devices

Weihua Cheng¹, Zhengyang Ding², Chao Xu¹*, Xiaohu Wu¹, Yuanyi Xia² and Jiaming Mao²

¹ Jiangsu Electric Power Information Technology Co., Ltd., China
² State Grid Jiangsu Electric Power Co., Ltd. information and communication branch

*dayangxx@163.com

Abstract. In recent years, with the rapid development of Internet of Things (IoT) technology, a large number of Internet of things devices such as network printers, webcams and routers have emerged in the cyberspace. However, the situation of network security is increasingly serious. Large-scale network attacks launched by terminal devices connected to the Internet occur frequently, causing a series of adverse effects such as information leakage and property loss to people. The establishment of a set of fingerprint generation system for Internet of things devices to accurately identify the device type is of great significance for the unified security control of the Internet of things. We proposed a RAFM which is a detection and identification system of IoT. RAFM consists two major module including auto detection and fingerprinting. RAFM collects messages sent by different Internet of things devices by means of passive listening. Based on the differences in the header fields of different devices, it USES a series of multi-class classification algorithms to identify device types. Simulation experiments show that RAFM can achieve an average prediction accuracy of 93.75%.

1. Introduction

Gartner [1] reports that as of 2016, there are over 6.4 billion IoT devices online worldwide, and it is estimated that the number of IoT devices worldwide will exceed 20 billion by 2020. According to the report of IoT Analytics[2], it is predicted that the number of Internet connections of devices in the world will maintain an average annual growth rate of 17% from 2019 to 2025, with the number of devices connected to cutting-edge 5G networks maintaining a compound annual growth rate of 113%.

We found that although the tens of thousands of companies are actively to provide various types of users of the Internet of things, but as a result of many traditional equipment manufacturers in the production of equipment does not take into account the factors of network security, computer security, led to the current spread in a variety of household environment, industrial production environment of IoT equipment has a great potential safety problems, such equipment vulnerable to hackers and other criminals implementation of network attack, which cause a harm to the whole network space.

In recent years, the form of network security is increasingly severe, criminals using the network terminal launched by the network attack time is also frequent, and the damage caused by the attack and the scope of influence has an obvious trend of expansion. In October 2016, Dyn was hit by a distributed denial of service (DDoS) attack. The attackers used Mirai virus to illegally access IoT devices found in the network with a large number of initial user passwords, and then continued to launch similar attacks on other devices they found after the breach. Attackers used Mirai virus to illegally control more than 1
million webcams, DVRS and other terminal devices, resulting in a large number of websites using Dyn DNS services unable to access the normal, including Twitter, GitHub, Airbnb, PayPal and other well-known Internet services crashed, leading to massive network chaos in the United States.

Such attacks "botnets" (botnets) again into people's horizons, Botnet is refers to some malicious virus software IoT device composed of network, so network can let hackers remotely DDoS attacks, a variety of networking equipment used in spam bombing, stealing sensitive key or blackmail software. The rapid development of Internet of things technology in recent years gives these hackers an opportunity. Internet of things based botnets are different from ordinary zombie viruses spread on the Windows platform. The connection between personal computers and servers will be protected by various anti-virus software and firewalls, while most of the Internet of things devices do not have such a security mechanism, and the number of devices is very large. Once controlled by hackers, Internet of things devices can not only spread zombie viruses more rapidly in a short period of time, but also cause a lot of adverse effects on social order. Some Internet of things viruses can invade traffic light systems, destroy various infrastructure in cities and cause social unrest. These cases show that the current world cyber attacks are extending to IoT devices, and IoT terminal security has become an indispensable part of the field of network security. It is imminent to establish and improve the security protection mechanism for these devices.

In our work, we introduce RAFM (a real-time auto-detecting and fingerprinting method for online IoT devices), a detection and identification system of IoT. In RAFM, we first detect real-time data of IoT device. The network protocol data of the device is captured, then we extract network protocol information and business information related to the terminal. On this basis, the RAFM discovers new IoT device in real time. Secondly, according to the collected information, the device fingerprint is used to identify the type of connected Internet of things terminal device. We use machine learning multi-classification recognition method to identify IoT device. Simulation experiments show that RAFM can achieve an average prediction accuracy of 93.75%.

The remainder of this paper is structured as follows. We begin by overviewing the related work on device fingerprint in Section 2. Next, in Section 3 we introduce RAFM. Further, in Section 4, we present the performance evaluation of RAFM, and then conclude the paper in Section 5.

2. Related Work

The discussion and research on the security of the Internet of things is no longer a new topic. With the occurrence of large-scale cyber attacks using the Internet of things devices in recent years, this topic has gradually become a hot topic for scholars at home and abroad.

Madaglia [3] proposed that with the explosive growth of the number of smart devices, the privacy protection mechanism of the IoT would become a major issue in the field of information security, especially the security of mobile Internet of things devices. To realize the security protection of interconnected intelligent devices, the most important work is to improve their identity authentication and management system.

Weber[4] proposed that to protect the security of the Internet of things, four aspects should be considered: preventing network attacks, data authentication, access control and protecting the privacy of users (natural and legal persons), as well as the fact that the sensor layer of the Internet of things cannot add advanced encryption protection algorithms due to limited hardware resources.

In recent years, in the research on security protection measures of Internet of things devices, device identification and identity authentication have always been the primary work. How to generate highly recognizable device fingerprint for Internet of things devices has gradually become a research hotspot.

In 2005, Tadayoshi [5] proposed a scheme to generate fingerprints for remote physical devices without the active help of the detected device. They realized the recognition of the device by using the small error of clock offset in part of the device's hardware.

In 2010, Cui [6] designed a scanning analysis system for embedded network devices, conducting extensive scanning of publicly accessible embedded network devices in cyberspace. The results showed that more than 540,000 embedded devices were still using the initial password set by the supplier, posing
a high security risk. These devices account for 13% of all scannable devices, including firewalls, routers and network printers used by businesses and VoIP adapters and IPTV set-top boxes used by home users. Cui et al. used the Nmap to perform network port scanning, and realized the identification of the above device types by using the initial outputs of Telnet and HTTP servers as features.

2015, Radhakrishnan [7] to design a kind of active and passive two ways may be adopted for the wireless local area network equipment GTID device type identification system, they used the same clock drift of the differences of different equipment hardware module, combined with statistical techniques using the network traffic data generation device type fingerprints, to join the artificial neural network classifier training, and use all the commonly used, on campus, tablets, smart phones and other devices tested a lot of experiments, the experimental results show that good accuracy.

In 2016, Cao [8] proposed a scheme to generate fingerprints by using HTTP data header data, and to use k-means algorithm for intelligent instrument authentication, and to identify Internet of things devices such as webcams, switches and routers with high accuracy. In another study, they proposed a method to extract the Web fingerprint of the network terminal device by using the fields of the HTTP data header and part of the information in the HTML source Web service.

In 2017, Ren [9] proposed a method to extract the characteristics of specific types of devices from the Web management pages of a large number of Internet of things devices, and use positive sample feedback to enhance the PU algorithm (FE-PU) to filter out the types of Internet of things devices that meet the requirements. An experiment was carried out on one million Web management pages in the network space to realize the recognition of the type of video surveillance equipment. That same year, Meidan [10] applying machine learning algorithms to connect to the Internet of things device recognition, through traffic information extracted from different Internet equipment, using supervised learning of multi-stage classifier for training, not only realized the IoT equipment and distinguish the content networking equipment, and to achieve the classification of Internet equipment, precision can reach 99.281%.

In 2018, Bezawada [11] also used the traffic data of Internet of things devices to extract features, realizing the generation of behavioral fingerprint of devices, and realizing the identification of devices of Internet of things with 99% accuracy. In the same year, Shaikh [12] proposed the use of machine learning classification algorithms, such as random forest algorithm and gradient enhancement algorithm, to identify infected Internet of things devices, thus providing enterprises with better tools to detect network system intrusion.

3. RAFM

The design of RAFM consists of two major components: Auto Detection and Fingerprinting.

3.1. Auto Detection

The Auto Detection is a device discovery module based cosine similarity match methods. During Auto Detection, we extract features to discover new IoT device. Features can be divided into two major categories: inherent features and protocol features. Inherent features represent characteristics of the device itself, such as device manufacturer, equipment specific parameters and other information. Protocol features represent the protocol characteristics of IoT device such as source IP address, destination IP address, source port ID, destination port ID, number of occurrences of a default eigenvalue in the transport layer protocol payload. We use \( v_1, v_2, v_3, \ldots, v_i \) for inherent features and \( v_{i+1}, v_{i+2}, v_{i+3}, \ldots, v_n \) for protocol features. We define a vector \( V \) to represent an unknown IoT device:

\[
V = \{v_1, v_2, v_3, \ldots, v_i\}
\]

Then we define a vector \( S_j \) to represent the eigenvectors of the sample data:

\[
S_j = \{s_{j1}, s_{j2}, s_{j3}, \ldots, s_{jm}\}
\]

Where \( j \) means \( j \)th type of known IoT device.
After RAFM detects an unknown device, the $V$ of the device is extracted and we use the cosine-similarity to judge whether the unknown device is an IoT of new type:

$$cosine \similarity = \frac{1}{\sqrt{\sum_{i=1}^{m} V_i^2} \sqrt{\sum_{j=1}^{n} S_j^2}} \sum_{m=1}^{m} \sum_{j=1}^{n} V_i S_j$$

(3)

If the calculated similarity is less than the pre-built threshold value, the unknown device is considered to be a new device, and the network traffic characteristics of the new device are stored in the database.

3.2. Fingerprinting

We collected the flow of IoT at startup for analysis. During the startup period, the device needs to switch from the power off state to the normal working state, and its main job is to load the relevant software and hardware configuration. This process has a strict order of work due to the dependencies between the various add-ons.

The network protocol framework of Internet of things devices consists of application layer, transmission layer and network layer. However, due to the wide variety of application layer protocols and relatively mixed data volume, we only select the transport layer and network layer for fingerprinting.

In the network layer, we mainly study the IP field of its header, and investigate whether there are differences in the IP field among different IoT devices. Available fields for IP headers include Version, Header Length, Differentiated Services Field, Total Length, Flags, Time to live, and Protocol. We find that many of the fields have a certain degree of correlation with the device itself, so the same fields collected on different devices will have significant differences.

- The version field occupies the size of 4bit, which describes the IP protocol version of the message. If the version is IPv4, the value of the field is 0100B, while if the version is IPv6, the value is 0110.
- In Differentiated Services Field, the first 6 bits are DSCP fields, and the last 2 bits are CU fields. DSCP is the distinguishing service code point, also known as DS marker value. These 8 bits combine with different codes to realize the control of different forwarding modes. The low 2 bit CU field displays the explicit congestion notification (ECN) field, which enables the big-ip system to proactively send the signal that the scheduling router is about to be overloaded to similar devices, thus ensuring that they can take timely countermeasures.
- The flag bit field size is 3bit, and only the last two bits are valid bits. The second bit is the Don’t fragment flag (DF). If this flag is 1, then the router cannot implement sharding on the current message after receiving it. The third bit is More fragments (MF), if a message through long need to shard, so in divided into multiple segments, we make fragments at the end of the logo is 0, said the current message for the last segment, the remaining fragments of the sign is 1, the receiver continually receiving MF of 1 until met MF s stop every block.
- Survival time segment has a 8 bit size, says the number of the datagram can be sent between different devices, is the message before being discarded allows for maximum network jump points, if the router receives the IP datagram TTL value is 0 or 1, the router will not forward the newspaper article, but will be discarded the datagram and sends the ICMP message overtime.

In the transport layer, we divide the transport layer protocols into three types for coding, including TCP, UDP and other protocols. The fields that can be used to extract features are PORT field, Maximum Segment Size, Window Size Value and Options.

- For the PORT field, according to the classification rules of IANA, the ports of the three categories of known PORT (0~1023), registered PORT (1024~49151) and dynamic PORT (49152~65535) are encoded as 0, 1 and 2 respectively.

The maximum segment length field is the limit of information that a message can send at one time. Before communicating with the receiver, TCP requires both sides to determine the maximum amount
of data they can receive at a single time, that is, the value of MSS field. After the maximum value is determined, if the amount of data sent by the sender exceeds the MSS limit, it must be sent in multiple fragments. The value of this field is usually equal to the length of the maximum transmission unit minus the network layer header length and the TCP header length.

According to the above analysis, we can generate a feature matrix $M$:

$$
M = \begin{bmatrix}
    f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,m} \\
    f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,m} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    f_{n,1} & f_{n,2} & f_{n,3} & \cdots & f_{n,m}
\end{bmatrix}
$$

(4)

Where $n$ represents the number of features adopted and $m$ represents number of packets. Then we can convert $M$ into a $m \times n$-dimensional vector $F$:

$$
F = (f_{1,1}, f_{1,2}, \ldots, f_{1,n}, f_{2,1}, f_{2,2}, \ldots, f_{2,n}, \ldots, f_{m,1}, f_{m,2}, \ldots, f_{m,n})
$$

(5)

Based on the above feature extraction process, we can obtain the final feature vector $F$:

$$
F = (F', Y)
$$

(6)

Where $Y$ represents the device type label for the Internet of things.

Then machine learning is applied to classify device fingerprints. Considering the simplicity and generalization, random forest is first adopted to our system. The classification model is an offline learning architecture. We collected a batch of traffic samples of setup stage for learning. Then we conduct the same experiment with SVM and logistic algorithm. The detailed performance report is given in section 4.

4. Performance Evaluation

In this Section, we will test the recognition effect of RAFM and analyze the results. We take each pcap file as a collection sample, propose the respective feature vectors from 10 devices, and combine the feature matrix as the input of the classification algorithm. We use BP neural network, support vector machine and k-Nearest Neighbor algorithm to test the above data.

We performed 10 cross-validations on the training results of the three classification models, that is, the data set was randomly divided into 10 parts, using 9 of them as the training set and the remaining 1 as the test set, and 10 times were performed on the model. For testing, average the accuracy of 10 classifications as the final algorithm accuracy estimate. The results are shown in Table 1.

| Test times | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|------------|----|----|----|----|----|----|----|----|----|----|
| BP accuracy| 75 | 87.5 | 87.5 | 93.75 | 81.25 | 87.5 | 87.5 | 87.5 | 87.5 | 87.5 |
| SVM accuracy| 93.75 | 87.5 | 87.5 | 82.5 | 88.75 | 88.75 | 87.5 | 93.75 | 85 | 88.75 |
| KNN accuracy| 12.5 | 12.5 | 12.5 | 18.75 | 12.5 | 18.75 | 18.75 | 6.25 | 12.5 | 6.25 |

It can be seen from the above results that the BP neural network and SVM algorithm obviously have good recognition effect for the feature mode adopted by this system. Compared with BP algorithm, SVM has a slight advantage. We choose to use SVM algorithm in RAFM.

We built a multi-classifier on SVM and each device type are trained with a binary classifiers. The identification accuracy for SVM is shown in Figure 1 and we find 8 devices’ accuracy ratio reaches over 90%. While the other two devices had an accuracy of only 80%. This may prove that RAFM is very unstable. But when we came to their confusion matrix which is been shown in Figure 2. The reason why the other two devices had an accuracy of only 80% is that they are similar device from TP-Link. That, in turn, proves RAFM's accuracy.
Figure 1. Identification ratio of SVM with 10 device types.

Figure 2. Confusion matrix of SVM with 10 device types.

5. Conclusion
In this paper, we proposed RAFM which is a detection and identification system of IoT. Experimental results show that RAFM is effective and the recognition accuracy can reach 93.75%. Not many types of IoT devices are tested in our work. In our future work, we will try to put RAFM on more application scenarios.

Acknowledgment
Project Name: Research and Application of Wireless Access Security Management Platform based on Device Fingerprint Technology
Project Source: Science and Technology Project of State Grid Jiangsu Electric Power Co., Ltd

References
[1] STANFORD C. Gartner says 6. 4 billion connected“things” will be in use in 2016, up 30 percent from 2015 [EB/OL]. [2017-1210]. http: www.gartner.com/newsroom/id/3165317.
[2] IoT Analytics. IoTAnalytics-State of the IoT Winter 2018-2019 -vf. http:iot-analytics.com/
[3] Medaglia C M, Serbanati A. An overview of privacy and security issues in the internet of things[M]/The internet of things. Springer, New York, NY, 2010: 389-395.
[4] Weber R H. Internet of Things–New security and privacy challenges[J]. Computer law & security review, 2010, 26(1): 23-30.

[5] Kohno T, Broido A, Claffy K C. Remote physical device fingerprinting[J]. IEEE Transactions on Dependable and Secure Computing, 2005, 2(2): 93-108.

[6] Cui A, Stolfo S J. A quantitative analysis of the insecurity of embedded network devices: results of a wide-area scan[C]//Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010: 97-106.

[7] Radhakrishnan S V, Uluagac A S, Beyah R. GTID: A technique for physical device and device type fingerprinting[J]. IEEE Transactions on Dependable and Secure Computing, 2014, 12(5): 519-532.

[8] CAO Laicheng, ZHAO Jianjun, CUI Xiang, LI Ke. Cyberspace device identification based on K-means with cosine distance measure[J]. Journal of University of Chinese Academy of Sciences, 2016, 33(4): 562-569.

[9] R.L. Ren, Y. Gu, J. Cui, S. Liu, H.S. Zhu, and L.M. Sun, “Web Features-based Recognition Specific-Type IoT Device in Cyberspace,” Communications Technology, vol. 50, no. 5, pp. 1003-1009 (in Chinese), 2017.

[10] Y. Meidan, B. Michael, S. Asaf, J.D. Guarnizo, O. Martin, N.O. Tippenhauer, and E. Yuval, “ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis,” In Proceedings of the Symposium on Applied Computing, pp. 506-509. April, 2017.

[11] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, “IoTSense: Behavioral Fingerprinting of IoT Devices,” arXiv preprint arXiv:1804.03852, Apr. 2018.

[12] Y. Meidan, B. Michael, S. Asaf, J.D. Guarnizo, O. Martin, N.O. Tippenhauer, and E. Yuval, “ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis,” In Proceedings of the Symposium on Applied Computing, pp. 506-509. April, 2017.