

A Device Identification Method Based on Combination of Multiple Information

Yuki Imamura
Graduate School of Engineering
Osaka City University

Nobuyuki Nakamura
Oki Electric Industry Co.,Ltd.

Taketsugu Yao
Oki Electric Industry Co.,Ltd.

Shingo Ata
Graduate School of Engineering
Osaka City University

Ikuo Oka
Graduate School of Engineering
Osaka City University

Abstract—Recently, security issues have become more serious with the spread of IoT devices. To reduce the security risks when using communication devices, network administrators should know detailed information of all devices connected in the network and control them accordingly. Although there are many studies about identifying devices based on information from the network, there is a limit to improving the accuracy of device identification by using a single identification method.

In this paper, we propose a new method to improve the accuracy of device identification by using a combination of results from multiple identification methods. We define a new index, called *similarity*, between multiple devices by using the results from multiple identification methods and identify devices based on the *similarity*. The results of evaluations for communication devices connected to a real network show that our method can improve the accuracy of device identification by 23.0 percent.

Index Terms—device identification, clustering, multiple information, similarity of devices, IoT

I. INTRODUCTION

With recent developments in communication technology, many kinds of communication devices are connected to a network, and the security of these communication devices has becomes a serious issue [1]. As the number of communication devices increases, the number of communication devices affected by security risks also increases. Moreover, IoT (Internet of things) devices have the capability to communicate over the network these days, and IoT devices are expected to become increasingly popular in the near future [2].

To reduce the security risks of communication devices, it is important to identify detailed information of all devices and control such devices by effective monitoring in a network performed by the network administrator. However, it is difficult for the network administrator to identify detailed information of all devices and control such devices in a network because the administrator should identify devices by only monitoring traffic between devices (i.e., not deploying monitoring software in devices). Currently, there are many methods of identifying a communication device based on information obtained by monitoring the communications of the device.

Recent studies on device identification have mostly been based on port numbers or traffic patterns [3]. However, a weak point of using port numbers as the identifier of devices is an end user can easily change the values. On the other hand, in most methods based on measurements of traffic patterns, traffic statistics such as packet size or packet inter-arrival time are used [4]. However, to identify devices accurately by using traffic patterns, a sufficient amount of traffic patterns should be collected in prior, which is used as training data for learning-based identification methods [5]. In this way, by using traditional identification methods based on individual information, we cannot identify many kinds of devices with high accuracy by using only a single identification method. Note that individual information means just one kind of information from the communication of each device.

We should deal with the problem by using multiple methods of identification of devices. Multiple information means more than two kinds of individual information whose sources are different from each other. If there are some devices we cannot identify by using a method based on individual information, we should identify the devices by using some methods based on the other information.

Note that we do NOT propose an identification method using *multiple information* but propose a method that combines the results of *multiple identification methods*. More specifically, many identification methods are categorized as classification algorithms, where a set of metrics is used.

The former method (i.e., the method using multiple information) simply uses a single classification algorithm by concatenating multiple vectors into a single vector. However, it is difficult to maximize the strength of each individual piece of information for two reasons: the characteristics of vectors from each piece of information are different from each other, and the optimal identification algorithm depends on the distribution of vectors.

For this reason, we focus on the latter method, i.e., we first obtain multiple results by using multiple identification methods and then use an additional clustering method to combine the multiple results to improve the accuracy. Thus, we identify devices based on the metric called the *similarity*

between devices, which is obtained from multiple identification results. By identifying devices based on the *similarity* between devices, even if some devices are not identified by a single identification method, they are identified by other identification methods, and that enables the identification of many kinds of devices with high accuracy.

As an evaluation example, we combine six identification methods in this paper. Then, we define the *identification accuracy* for evaluation. As a result of the evaluation for communication devices connected to the real network, we can improve the *identification accuracy* by 23.0 percent.

The rest of this paper is organized as follows. Section II surveys related work, and Section III describes our device identification algorithm based on the *similarity* between devices. In Section IV, we present the experimental results. Finally, Section V concludes the paper and presents possible future work.

II. RELATED WORKS

There are many studies focusing on communication device identification based on several kinds of information such as port numbers, traffic patterns [6], or payload analysis [7]. In particular, studies on device identification based on traffic patterns have been proposed a lot [8].

In [9], the authors use a deep learning algorithm, where they use only the packet inter-arrival time of received packets as features and identify the iPad 4 and iPhone 7 Plus¹ with an accuracy of 86.7%. They experiment on a network that is connected to only two kinds of devices. Based on only the packet inter-arrival time of received packets, it may be difficult to identify many kinds of devices with high accuracy in a real network, where much of the traffic is mixed.

In [10], the authors use distributions of packet sizes and packet inter-arrival times as identification features. They identify nine kinds of devices including the same kinds of devices made by different vendors with an accuracy of 88.3% on average by using the statistics and a support vector machine algorithm. However, the accuracy of identification of some kinds of devices is 69.8%, and the accuracy is far worse than the overall accuracy of device identification. They report that it is hard to get enough traffic for identification because such kinds of devices generate less traffic to communicate.

In this way, it is not easy to identify many kinds of devices with high accuracy by using conventional identification methods based on individual information. Thus, in this paper, we propose a method to identify many kinds of devices with high accuracy based on a combination of results from multiple identification methods and show that the overall accuracy of identification by using our method is higher than that by using any methods based on individual information.

III. IDENTIFICATION METHOD

We show the procedure of our identification method in Fig. 1. In this method, we identify R devices by combining

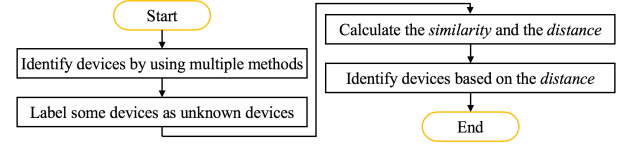


Fig. 1. Procedure of Proposed Method

N methods. The proposed device identification algorithm is as follows.

- 1) Identify R devices by using N methods.
 - a) Identify R devices by using a *primary method*
 - b) Cluster R devices by using $N - 1$ *supplement methods*
- 2) Extract R' devices that satisfy both of the following two requirements:
 - Device that we can get information from in either of N methods.
 - Device that is not *unknown* in either of N methods.
- 3) Calculate the *similarity* and the *distance* between multiple devices.
- 4) Cluster devices based on the *distance* between multiple devices.

In the following subsections, we explain every step in Fig. 1 separately.

A. Identifying Devices by Using Multiple Methods

In this paper, as a specific way to improve the overall accuracy of device identification based on a combination of results from multiple identification methods, we propose a way to supplement a certain identification result with the other identification results. We define a method to be supplemented as a *primary method* and methods to supplement the *primary method* as *supplement methods*.

A device identification algorithm mainly consists of the following two steps: a step to get clusters that consist of the same kind of device (defined as a *clustering step*) and a step to label each cluster as a kind of device (defined as a *labeling step*). Since the *supplement methods* are used for supplementing the *primary method*, we use clusters from the *clustering step* in the *supplement methods* as identification results to supplement the *primary method*.

In this paper, we combine six identification methods (named I_1 to I_6).

B. Labeling Some Devices as Unknown Devices

When we cannot identify some devices by using any methods based on individual information, we cannot identify the devices by combining the methods. Thus, we label devices not identified by using any methods based on individual information as *unknown* devices.

C. Calculating the Similarity and the Distance

To judge whether multiple devices are the same kind of device quantitatively, we define an index that shows how similar multiple devices are from the multiple identification results, called *similarity*, between multiple devices.

¹Apple, iPad, and iPhone are registered trademarks of Apple Inc.

The simplest solution of the *similarity* between multiple devices is the number of identification methods that identify these devices as the same kind of device. In this case, however, the *similarity* of different kinds of device may be higher than that of the same kinds of device because considering all identification methods has the same level of importance for device identification. The importance of each identification method is different from each other due to the following three factors: *granularity of clusters*, *accuracy of clusters*, and *correlation with other methods*.

To determine the *similarity* based on the characteristics of each method, we define an index that shows the level of importance for device identification, called the *weight* of the identification method. Then, we consider that an identification method with a large *weight* is more important than that with a small *weight* for device identification. Subsequently, we make it easier for the results from identification methods with a large *weight* to be reflected in the final identification result than those with a small *weight*. Since the *weights* of each identification method vary depending on each method and whether we can get information, we define two indices A_n and B_n as the *weights* of the n th identification method. If we can get the information from a source and identify devices based on the information, the *weight* of the result is A_n . The *weight* of the result when we cannot get the information from a source is B_n . By defining the *similarity* between multiple devices based on A_n and B_n , the identification result of our method makes it easier for the results from identification methods with high importance to be reflected than those with low importance.

1) Weight of Each Identification Method:

a) *Granularity of Clusters*: All identification methods based on individual information are composed of methods that have various meanings such as a pair of the kind of device and vendor, the kind of device, and the operating system running on the device. To make our algorithm generic, we should use objective indices because the meanings are subjective.

Therefore, we use the *granularity of clusters* as the index evaluating the meanings of clusters. When we identify devices by using an identification method with a large *granularity of clusters*, different kinds of devices are likely to be identified as the same kind of devices. Then, a method with a small *granularity of clusters* is more important than that with a large *granularity of clusters*.

b) *Accuracy of Clusters*: The accuracy of each identification method is different from each other. Identification methods with low accuracy means that different kinds of devices are likely to get the same identification result from the identification methods. Then, a method with high accuracy is more important than that with low accuracy. However, we cannot evaluate the accuracy of each identification method because we do not know the kinds of device when we try to identify the devices.

When different kinds of devices are identified as the same kind of devices, the causes are errors in getting information and inaccuracy in identifying. The errors in getting information

cannot be evaluated with statistics got from the identification process. On the other hand, the accuracy in identifying can be evaluated with the distance between clusters. If the distance between clusters from an identification method is far, the clusters are discrete. The fact that the clusters are discrete means that the distance between vectors representing different kinds of devices is far. If clusters are discrete, identification results tend to be accurate because the features of different kinds of device differ greatly.

c) *Correlation with Other Methods*: If some sources of information are similar, the information from the sources strongly affects the identification results of our proposed method because multiple identification results from the methods based on the sources are also similar.

Regarding indices to evaluate the similarity of multiple information objectively, sets of devices whose information can be got from similar sources are also likely to be similar. Then, we use the overlap coefficient to increase the effect of the number of common elements. There is the problem that the overlap coefficient between a set and a subset of the set is 1, which is the maximum value of the overlap coefficient because of not considering the difference in the number of elements. Thus, we define the *correlation with other methods* with considering the difference in the number of elements.

d) *Weight of the Identification Method*: We define the *weight* of the identification method by the *granularity of clusters*, the *accuracy of clusters* and *correlation with other methods*. When the *granularity of clusters* or the *accuracy of clusters* becomes larger, the *weight* also becomes larger because the method is important. On the other hand, when the *correlation with other methods* becomes larger, the *weight* becomes smaller.

2) *Definition of the Similarity and the Distance*: When we combine multiple identification methods, important methods should affect the identification result from the method combining multiple identification methods. Thus, we define the *similarity* between multiple devices as Eq. (1).

$$s_{i,j} = \sum_{n=1}^N u_{i,j}(n) \quad (1)$$

$$u_{i,j}(n) = \begin{cases} A_n & (L^n(M_i) = L^n(M_j) \neq \text{None}) \\ B_n & (L^n(M_i) = L^n(M_j) = \text{None}) \\ 0 & (\text{other}) \end{cases}$$

Where M_r is the r th device, $L^n(M_r)$ is M_r 's identification result from n th method, and $s_{i,j}$ is the *similarity* between M_i and M_j .

Next, we define the *distance* between multiple devices by the *similarity*. If the *similarity* becomes larger, the *distance* between multiple devices should become smaller.

D. Identifying Devices Based on the Distance

We identify devices based on the *distance* between multiple devices by using a clustering algorithm. In this paper, we use Ward's method as a hierarchical clustering algorithm, since the method is known to be less sensitive to exceptional samples and capable of accurately sorting samples.

TABLE I
DEVICES WE IDENTIFY IN THIS PAPER.

Kind of device	Symbol	Number of devices
PC or server	X_1	44
Wireless LAN access point	X_2	11
Network attached storage (NAS)	X_3	4
Printer	X_4	12
Switching hub	X_5	3

Finally, we decide the number of clusters. Since the identification accuracy depends on the number of clusters, we should use the number of clusters that maximizes the identification accuracy as the optimal number of clusters to identify devices. However, we cannot evaluate the identification accuracy when we decide the number of clusters because we do not know the kinds of device when we try to identify devices. To decide the optimal number of clusters, some indices for evaluating clusters are proposed. There are mainly indices for evaluating the discreteness between clusters, the cohesiveness within clusters, or both. We should use indices that evaluate both the discreteness between clusters and the cohesiveness within the clusters when we use the index to decide the number of clusters. Thus, we decide the optimal number of clusters by using Pseudo F [11].

IV. IDENTIFICATION RESULT

We show the devices to be identified in this paper in Table I. The devices are actually connected to the network and operated. We show the identification accuracy of each individual identification method and our proposed method in Table II. The overall identification accuracy of our method is higher than that of each individual method.

Focusing on each kind of device, the identification accuracy on a PC or server of our method is higher than that of each individual method. There are multiple methods to identify the PC or server, and we can identify different kinds of PC or server by each method. Then, we can identify the PCs or servers that are difficult to identify by individual information by using the results from other identification methods.

The identification accuracy on the wireless local area network (LAN) access point of our method is as high as the maximum identification accuracy on the wireless LAN access point of each individual method. There are also multiple methods to identify the wireless LAN access points. However, we cannot identify some wireless LAN access points by all of the multiple methods. As a result, the identification accuracy on the wireless LAN access points of our method is not higher than that of all of the multiple methods, like the identification accuracy on a PC or server.

The identification accuracy on the NAS and printer of our method is also as high as the maximum identification accuracy on the NAS of each individual method. There is only one method to identify the NAS and printer in multiple methods.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a method to improve the overall identification accuracy of device identification based on a combination of results from multiple identification methods.

TABLE II
IDENTIFICATION ACCURACY OF EACH INDIVIDUAL IDENTIFICATION METHOD AND OUR PROPOSED METHOD.

	I_1	I_2	I_3	I_4	I_5	I_6	Our method
X_1	0.0%	79.5%	81.8%	25.0%	9.1%	6.8%	100.0%
X_2	54.5%	54.5%	36.4%	0.0%	0.0%	0.0%	54.5%
X_3	0.0%	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
X_4	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%
X_5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Overall	13.5%	55.4%	54.1%	14.9%	8.1%	6.8%	78.4%

We defined the weight of the identification method and the *similarity* between multiple devices based on the weight of the identification method. As a result, the overall identification accuracy of our proposed device identification method is higher than that of each individual identification method.

For future research, we should increase the number of devices to identify and prove that our method improves device identification accuracy in any environment.

ACKNOWLEDGEMENTS

This study is supported by Grants-in-Aid for Scientific Research of Japan (19H04097). We are grateful to Prof. Yoshida Masamichi, Graduate School of Science, Osaka City University, and Mr. Kamae Tetsuro for helpful discussions.

REFERENCES

- [1] M. Abomhara *et al.*, "Cyber security and the internet of things: Vulnerabilities, threats, intruders and attacks," *Journal of Cyber Security and Mobility*, vol. 4, no. 1, pp. 65–88, 2015.
- [2] C.-L. Wu and L.-C. Fu, "Design and realization of a framework for human–system interaction in smart homes," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 1, pp. 15–31, 2011.
- [3] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, "Iotsense: Behavioral fingerprinting of iot devices," *arXiv preprint arXiv:1804.03852*, 2018.
- [4] M. R. Shahid, G. Blanc, Z. Zhang, and H. Debar, "Iot devices recognition through network traffic analysis," in *Proceedings of 2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 5187–5192.
- [5] M. N. Msadek, R. Soua, and T. Engel, "Iot device fingerprinting: Machine learning based encrypted traffic analysis," in *Proceedings of 2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019.
- [6] Y. Meidan, M. Bohadana, A. Shabtai, M. Ochoa, N. O. Tippenhauer, J. D. Guarnizo, and Y. Elovici, "Detection of unauthorized iot devices using machine learning techniques," *arXiv preprint arXiv:1709.04647*, 2017.
- [7] J. Xiao-jiao, Z. Cheng, Y. Liu, and L. Yong-nian, "Identifying the p2p flow based on traffic and payload characteristics," *Computer Engineering & Science*, vol. 8, 2010.
- [8] L. Bai, L. Yao, S. S. Kanhere, X. Wang, and Z. Yang, "Automatic device classification from network traffic streams of internet of things," in *Proceedings of 2018 IEEE 43rd Conference on Local Computer Networks (LCN)*. IEEE, 2018, pp. 1–9.
- [9] S. Aneja, N. Aneja, and M. S. Islam, "Iot device fingerprint using deep learning," in *Proceedings of 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*. IEEE, 2018, pp. 174–179.
- [10] H. Kawai, S. Ata, N. Nakamura, and I. Oka, "Identification of communication devices from analysis of traffic patterns," in *Proceedings of 2017 IEEE 13th International Conference on Network and Service Management (CNSM)*. IEEE, 2017, pp. 1–5.
- [11] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.