

Analysis of DNS Graph of Phishing Websites Using Digital Certificates

Yuki Ishida*, Masaki Hanada†, Atsushi Waseda†, Moo Wan Kim‡

* Graduate School of Informatics, Tokyo University of Information Sciences, Japan

† Department of Informatics, Tokyo University of Information Sciences, Japan

‡ TA Tech., Japan

Email: h22001iy@edu.tuis.ac.jp, mhanada@rsch.tuis.ac.jp

Abstract— The prevalence of phishing threats has grown substantially over the past several years. Because phishing domains generally have a very short lifetime to avoid detection, they must be identified as soon as possible to prevent damage. In general, phishers frequently reassign resources, such as domain names and IP addresses, to avoid detection. Consequently, multiple domains may be hosted by the same IP address, or vice versa. Furthermore, phishing websites often employ certificates and HTTPS encryption to appear more trustworthy. In this study, we focused on HTTPS-enabled phishing websites to construct and analyze DNS graphs of domain names and IP addresses of phishing websites using Certificate Transparency (CT) logs. From the analysis results, we examined the differences between benign and phishing websites in terms of the number of nodes per component and average node degree.

Keywords— Phishing, DNS, DNS Graph, Security

I. INTRODUCTION

Phishing attacks are currently on the rise, with the Anti-Phishing Work Group (APWG) recently announcing 1,097,811 such attacks observed over the second quarter of 2022. This is the highest figure since the APWG began observations [1].

One of the various approaches proposed to identify phishing websites is the block list method. However, this approach is highly time-intensive, and many phishing websites are already closed by the time they are block listed. Because the mean lifetime of a phishing website is three days, these websites must be detected quickly.

Recently, phishing websites have begun to adopt certificates and HTTPS encryption to appear more trustworthy. The Phishing Activity Trends Report published by the APWG claims that 82% of phishing websites reported in the first quarter of 2021 have used HTTPS [2]. HTTPS-encrypted websites use certificates to establish their identity and ensure secure web communication. These certificates must be obtained from trusted certification authorities (CAs). However, the recent emergence of free CAs, such as Let's Encrypt and cPanel, has allowed phishers to

easily and mechanically obtain certificates. To improve the transparency of these certificates, many free and commercial CAs use Certificate Transparency (CT) to verify their legitimacy. CT is a security framework for monitoring and auditing all public digital certificates issued by CAs [3], which stipulate that whenever a CA issues a digital certificate, an appropriate log entry is made. Using the CT logs, we can therefore analyze the behavior of phishing websites.

In the present study, we focused on phishing websites using HTTPS to construct and analyze DNS graphs of domain names and IP addresses corresponding to phishing websites using CT logs.

Phishers generally reassign resources – specifically domain names and IP addresses – to avoid detection. Accordingly, we constructed DNS graphs of domain names and IP addresses of phishing websites using CT logs, and analyzed the graph using complex network metrics such as the clustering coefficient and node degree. Specifically, we first fetched digital certificates from CT logs, and retrieved the corresponding domain names. Next, the obtained domain names were translated to IP addresses, which in turn were translated to domain names by DNS. Subsequently, we constructed DNS graphs to express the correspondence relationship between domain names and IP addresses, as well as categorized the data among phishing and benign websites. Finally, we compared the DNS graphs of phishing and benign websites using the clustering coefficient and node degree.

The remainder of this paper is organized as follows. Section II presents a summary of related work. Section III describes the DNS graph model employed in this study. Section IV discusses the construction process of the DNS graphs. Section V presents the analysis results for the DNS graphs using complex network metrics. Finally, Section VI concludes the paper.

II. RELATED WORKS

The block list method has been proposed as a typical approach for detecting phishing websites based on domain names and/or

IP addresses. This method lists the reported phishing domains and IP addresses in advance and blocks access to them. *Safe Browsing* [4] is a typical URL-based block list method provided by Google Inc. whereas *1.1.1.1 for Families* [5] is a DNS-based block list method provided by Cloudflare Inc.. However, these methods only block the domain names and/or IP addresses of known phishing websites. To avoid detection, phishers quickly generate and discard URLs, making it difficult for these methods to keep up. To address this issue, Prakash et al. examined current block lists and generated new URLs by employing various heuristics (e.g., changing the top-level domains). They proposed a method based on approximate matches with these generated URLs [6]. To detect phishing websites adopting HTTPS, Sakurai et al. created a regular expression template using domain name features of phishing websites from CT logs [7]. Subsequently, they demonstrated that certain phishing websites could be detected using this template.

Despite the aforementioned variety in block list approaches, the methods remain insufficient with regard to dealing with rapidly changing domain names and IP addresses.

III. DNS GRAPH MODEL

To avoid detection, phishers frequently change the domain names and IP addresses of websites. As a result, multiple domains may be hosted by the same IP address or vice versa. Furthermore, a possibility exists for phishers to attack using multiple domain names simultaneously. In this case, these domains are typically hosted by the same IP address.

In this study, we constructed and analyzed DNS graphs to assess these circumstances. In [8], the following procedure was employed to construct a DNS graph. By analyzing the *A records* in the DNS response, two columns (d, IP) were constructed, where d is the domain resolved to IP (IP address). The result is a bipartite graph that plots domains and IP addresses. The edge is formed between a domain d and an IP if there exists a record (d, IP). A similar structure was employed throughout the present study. Here, each component is a connected subgraph. Whereas the DNS graph in [8] was constructed using the *A records* in the DNS response, we constructed our graphs using the *A*, *CNAME*, and *PTR records* in the DNS response. We used the *CNAME records* because many domain name aliases may be hosted by the same IP addresses for malicious purposes. The *PTR records* were generally considered for security and verification. However, when phishers use a cloud service, the service provider may set the *PTR records* irrespective of the phisher's wishes. Using the resulting DNS graphs, we could analyze the frequent domain name and IP address changes of phishing websites.

A sample DNS graph is presented in Figure 1.

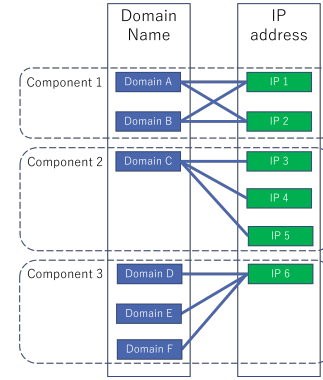


Figure 1. Example of DNS Graph

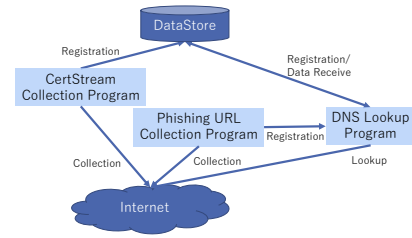


Figure 2. Proposed collection and analysis system

IV. DNS GRAPH BUILDING

To construct DNS graphs of phishing and benign websites, we developed a system to collect certificate information from CT logs and perform name resolution using DNS. This system was employed to compile the domain names of phishing websites from web sources. An overview of the data collection system is illustrated in Figure 2.

All the CT logs were collected by CertStream [9], a library used for receiving certificate list updates in real time.

The following procedure was employed in the collection of CT logs and construction of DNS graphs:

Step 1: The collection system first fetches digital certificates from CT logs via CertStream, and retrieves all website domain names from these digital certificates.

Step 2: A obtained domain name is translated to an IP address (i.e., value of the *A record* in the DNS response) via forward DNS lookup. In addition, if the *CNAME record* is included in the DNS response, the corresponding domain name (i.e., value of the *CNAME record*) is similarly translated to an IP address.

Step 3: The IP address obtained in **Step 2** is translated to a domain name (i.e., value of the *PTR record* in the DNS response) via reverse DNS lookup.

Step 1 is conducted daily. For each iteration of **Step 1**, **Steps 2** and **3** are repeated, the relationships between domain names and IP addresses are clarified, and a DNS graph is constructed accordingly. Here, **Steps 2** and **3** are alternately processed every other day.

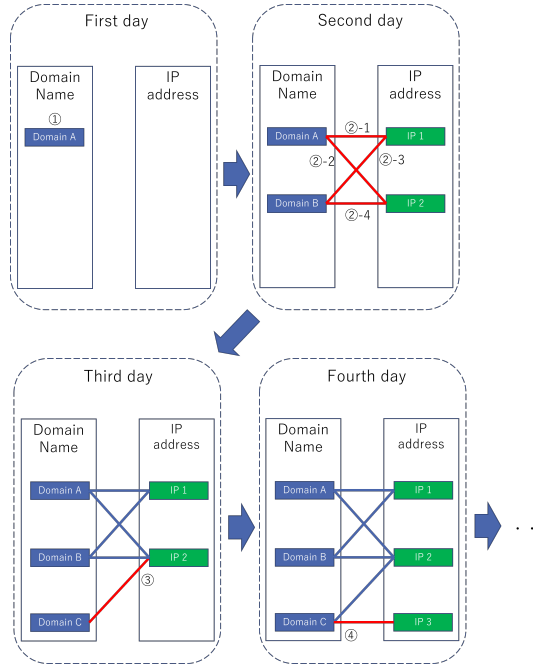


Figure 3. Example of the DNS graph building procedure

A sample run of this procedure is illustrated in Figure 3. Here, we assume that *Domain A* was retrieved from digital certificates of CT logs via **Step 1** on the first day (① in Figure 3). On the second day, *Domain A* was translated to *IP 1* and *IP 2* (i.e., the value of the *A record*) via forward DNS lookup. The relationship between *Domain A* and *IP 1* as well as that between *Domain A* and *IP 2* (i.e., $(DomainA, IP1)$ and $(DomainA, IP2)$) were then constructed as seen in ②-1 and ②-2 of Figure 3. In addition, assuming that *Domain B* (i.e., the value of the *CNAME record*) is included in the DNS response, the relationships between *Domain B* and *IP 1* and between *Domain B* and *IP 2* (i.e., $(DomainB, IP1)$ and $(DomainB, IP2)$) are simultaneously constructed as shown in ②-3 and ②-4 of Figure 3. On the third day (**Step 3**), assuming that *IP 2* can be only translated to *Domain C* (i.e., value of the *PTR record*) via reverse DNS lookup, the relationship between *Domain C* and *IP 2* (i.e., $(DomainC, IP2)$) is constructed according to the figure (③ of Figure 3). On the fourth day (**Step 2**), *Domain C* is translated to *IP 3* via forward DNS lookup. The relationship between *Domain C* and *IP 3* (i.e., $(DomainC, IP3)$) is then constructed (④ of Figure 3).

The aforementioned example focuses on *Domain A*, which is retrieved in **Step 1**. In fact, because **Step 1** is conducted every day, a lot of domain names are retrieved in **Step 1**. Therefore, each procedure after **Step 1** is conducted over a different series of days.

We then split the cumulative DNS graph into the phishing and benign subgraphs. For this purpose, we needed to collect

TABLE 1. Configuration results of cumulative DNS graph

Number of nodes (both benign and phishing)	45572
Number of domain names	36541
Number of IP addresses	9031
Number of components	6647

TABLE 2. Configuration results of DNS graph for benign websites

Number of nodes (benign)	4259
Number of domain names	2480
Number of IP addresses	1779
Number of components	1539

TABLE 3. Configuration results of DNS graph for phishing websites

Number of nodes (phishing)	4754
Number of domain names	4225
Number of IP addresses	529
Number of components	323

the domain names and IP addresses of phishing websites. We obtained this information from certain sites listing reported phishing URLs (i.e., PhishTank [10], OpenPhish [11], PhishStats [12] and Council of Anti-Phishing Japan [13]). In accordance with the aforementioned **Steps 2** and **3**, we then create a list of domain names or IP addresses of phishing websites. Finally, if a component of the constructed DNS graph includes an element of the list, it is labeled as a phishing website.

V. EXPERIMENT ANALYSIS

In the following section, we describe our experimental conditions, and present analysis results for the DNS graphs of phishing and benign websites.

A. Experimental Conditions

In this experiment, we collected domain names and IP addresses observed over a 7-day period from September 20, 2022 to September 27, 2022, and subsequently constructed DNS graphs. To reduce the load of DNS servers, the name resolution was conducted using randomly selected domain names and IP addresses. Table 1 lists the number of nodes, domain names, IP addresses, and components of the cumulative DNS graph, wherein each node represents a domain name or IP address. After splitting the graph into benign and phishing subgraphs, the aforementioned attributes of both subgraphs were listed in Tables 2 and 3 respectively.

B. Analyses of DNS graphs

We analyzed all DNS graphs using complex network metrics; specifically node degree, component size, and clustering coefficient.

First, we visualized the DNS graphs for benign and phishing websites, as shown in Figures 4 and 5 respectively.

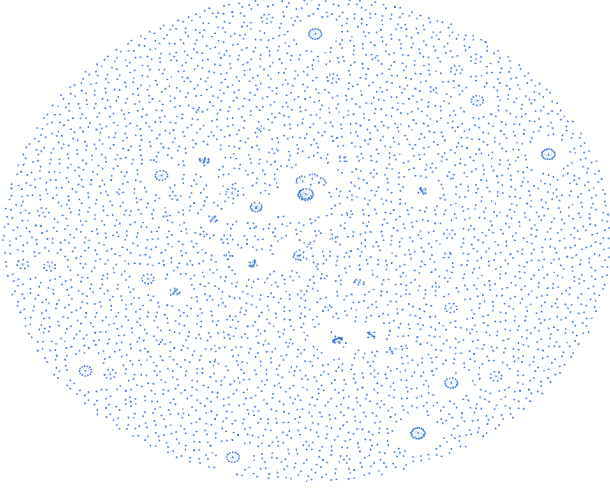


Figure 4. Visualization of DNS graph for benign websites

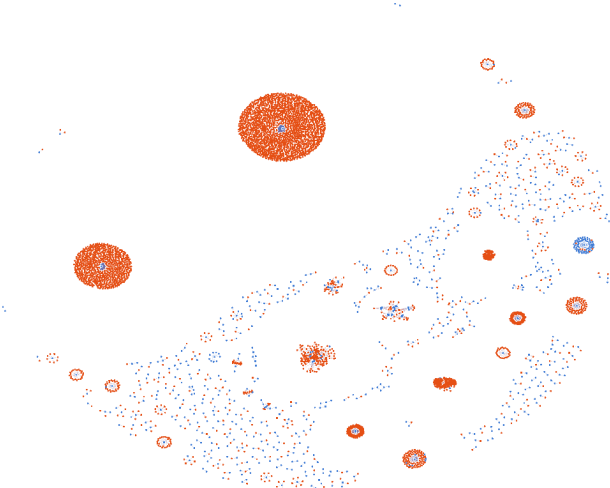


Figure 5. Visualization of DNS graph for phishing websites

The mean of the number of nodes per component, and that of the average node degree per component, for the benign and phishing websites are listed in Tables 4 and 5 respectively. Both figures are higher for phishing websites than for benign websites. Because a lot of combinations between domain name and IP address occur in components of phishing websites, it is assumed that phishers deploy attacks using multiple domain names and IP addresses simultaneously. Furthermore, multiple domains may be hosted by the same IP address or vice versa.

Certain components exhibit especially many nodes in the DNS graphs shown in Figure 4 and 5. Figure 6 presents a histogram measuring the number of nodes per component for benign websites, and Figure 7 presents an equivalent histogram for phishing websites. Phishing websites with components that encompass

TABLE 4. Analysis results of DNS graph for benign websites

Mean of the number of nodes per component	2.767
Mean of average node degree per component	1.186

TABLE 5. Analysis results of DNS graph for phishing websites

Mean of the number of nodes per component	14.718
Mean of average node degree per component	1.457

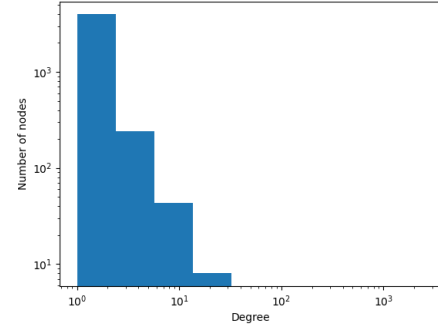


Figure 6. Histogram of average node degree for benign websites

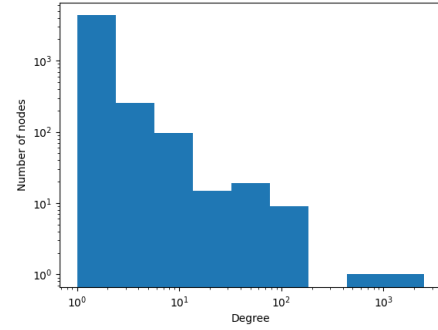


Figure 7. Histogram of average node degree for phishing websites

many nodes pose a greater threat to users. Accordingly, we extracted any components associated with 10 or more nodes from the DNS graphs for further analysis. Figure 8 presents a visualization of the DNS graphs for benign websites, and Figure 9 presents an equivalent visualization for phishing websites. Tables 6 and 7 list the analysis results for benign and phishing websites respectively.

We focus on components with 10 or more nodes of the DNS graphs of the benign and phishing websites and analyzed the cluster coefficient for the previously extracted components. In general, the triangle-based clustering coefficient measures how the neighboring nodes of a selected node are connected. How-

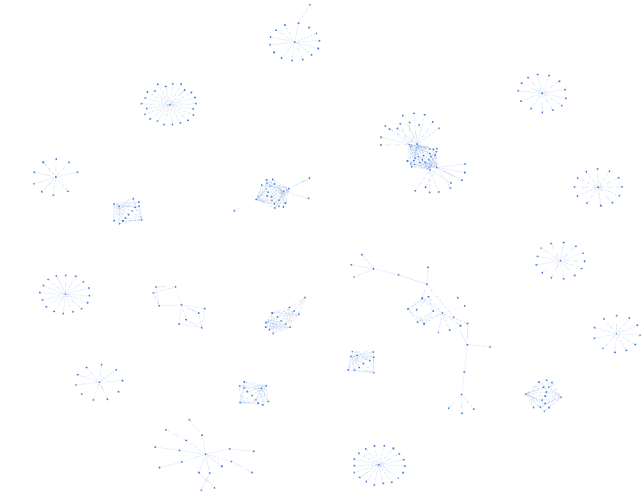


Figure 8. DNS graph of components with 10 or more nodes in benign websites

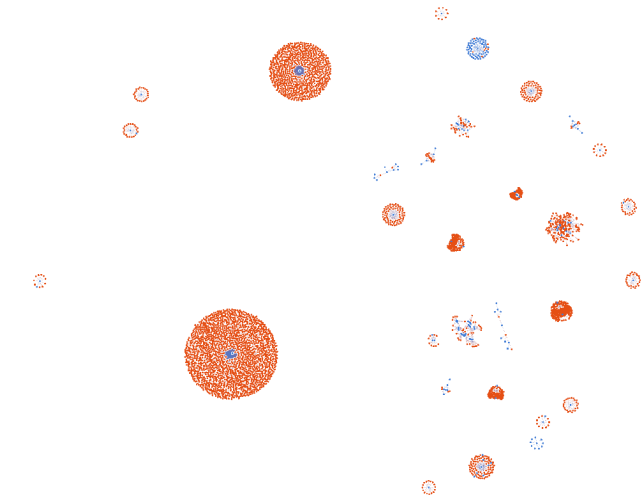


Figure 9. DNS graph of components with 10 or more nodes in phishing websites

ever, this coefficient cannot be applied in the present study, as the DNS graphs are bipartite. Instead, the square-based clustering coefficient proposed by Zhang et al. [14] was employed.

Figures 10 and 11 show the square-based clustering coefficients of each component (i.e., the x-axis is the numbered component). When the cluster coefficient is zero, it indicates star topology (hub network). 55% (11/20) of the selected components of benign websites exhibited star topology, whereas 60% (18/30) of the phishing websites did. Multiple domains were hosted by a single IP address in more than half of both sets, with the ratio being somewhat higher for phishing websites than benign websites. For benign websites, star topology is assumed to emerge as a consequence of public hosting services. In the case of phishing websites, however, it is assumed that phishers

TABLE 6. Analysis results for components with 10 or more nodes in benign websites

Number of nodes	338
Number of domain names	237
Number of IP addresses	101
Number of components	20
Number of nodes per component	16.9
Average node degree	3.012

TABLE 7. Analysis results for components with 10 or more nodes in phishing websites

Number of nodes	3811
Number of domain names	3701
Number of IP addresses	110
Number of components	30
Number of nodes per component	127.033
Average node degree	2.689

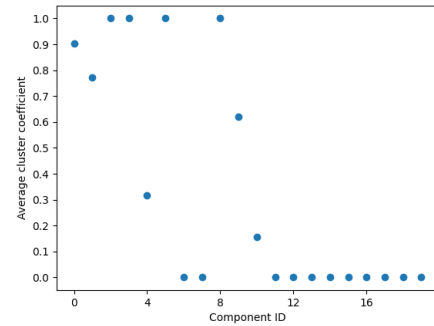


Figure 10. Cluster coefficients for benign websites with more than 10 nodes

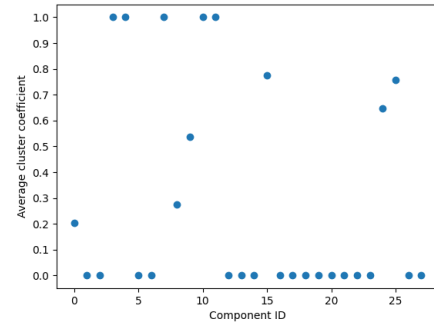


Figure 11. Cluster coefficients for phishing websites with more than 10 nodes

can attack using multiple domains in star topology. However, given the possibility to block access from multiple domains simultaneously via IP filtering, it may be possible to quickly identify the phishing websites. In the near future, we intend to conduct additional studies to achieve a sufficient understanding of this topic using time series analyses.

The mean of the square-based clustering coefficient for benign

websites is 0.751 when omitting any components with star topology, and for the corresponding figure for the phishing websites, it is 0.745. Regarding phishing websites with higher clustering coefficients, it is assumed that phishers generate multiple URLs and discard them after a certain period of time, before generating new URLs to avoid detection. The higher clustering coefficient exhibited by benign websites will be investigated as future work.

VI. CONCLUSION

In this study, we focused on phishing websites using HTTPS and constructed and analyzed DNS graphs of domain names and IP addresses of phishing websites using CT logs.

The main contributions are as follows: (1) DNS graphs of domain names and IP addresses belonging to HTTPS-enabled phishing websites were constructed and analyzed. (2) The primary differences between the DNS graphs of benign and phishing websites are the mean number of nodes per component, and that of the average node degree per component. (3) With regard to the clustering coefficient, few differences were observed between the benign and phishing websites corresponding to components with 10 or more nodes.

In the future, we plan to analyze the results of this study in more detail using other complex network metrics. In addition, we intend to analyze the growth and change of DNS graphs for benign and phishing websites using time series analysis, as well as develop a detection method for phishing websites.

REFERENCES

- [1] Anti-Phishing Working Group, "Phishing activity trends report 2nd quarter 2022," https://docs.apwg.org/reports/apwg_trends_report_q2_2022.pdf, (Accessed 14-Oct-2022).
- [2] Anti-Phishing Working Group, "Phishing activity trends report 2nd quarter 2021," https://docs.apwg.org/reports/apwg_trends_report_q2_2021.pdf, (Accessed 14-Oct-2022).
- [3] A. L. B. Laurie and E. Kasper, "Certificate transparency," *RFC*, vol. 6962, pp. 1–27, 2013.
- [4] "Google safe browsing," <https://developers.google.com/safe-browsing/>, (Accessed 14-Oct-2022).
- [5] "Introducing 1.1.1.1 for families," <https://blog.cloudflare.com/introducing-1-1-1-1-for-families/>, (Accessed 14-Oct-2022).
- [6] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: Predictive blacklisting to detect phishing attacks," in *2010 Proceedings IEEE INFOCOM*, 2010, pp. 1–5.
- [7] Y. Sakurai, T. Watanabe, T. Okuda, M. Akiyama, and T. Mori, "Identifying the phishing websites using the patterns of tls certificates," *Journal of Cyber Security and Mobility*, vol. 10, no. 2, pp. 451–486, 2021.
- [8] W. R. Futai Zou, Siyu Zhang and P. Yi, "Detecting malware based on dns graph mining," *International Journal of Distributed Sensor Networks*, vol. 2015, pp. 1–12, 2015.
- [9] "Certstream," <https://certstream.calidog.io/>, (Accessed 03-Oct-2022).
- [10] "Phishtank," <https://phishtank.org/>, (Accessed 05-Oct-2022).
- [11] "Openphish," <https://openphish.com/>, (Accessed 05-Oct-2022).
- [12] "Phishstats," <https://phishstats.info/>, (Accessed 05-Oct-2022).
- [13] "Council of anti-phishing japan," <https://www.antiphishing.jp/>, (Accessed 05-Oct-2022).
- [14] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, "Clustering coefficient and community structure of bipartite networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875, dec 2008.



Yuki Ishida received the B.E. degree and M.S. degrees in Informatics from Tokyo University of Information Sciences, Japan, in 2014 and 2016, respectively. He worked at Digital Arts, Inc in 2016, and had been engaged in developing security products. In 2019 he joined SecureBrain Corporation in Japan, and has been engaged in research and development for cyber security. At the same time, he has been enrolled at the doctoral program of Graduate School of Informatics, Tokyo University of Information Sciences in Japan since 2022. His research interests include cyber security and network quality control. He is a member of the IEEE, IEICE and IPSJ.



Masaki Hanada received the B.E. degree in resources engineering from Waseda University in 1996, the M.S. degree in information science from Japan Advanced Institute of Science and Technology (JAIST) in 1999, and the M.S. and D.S. degrees in global information and telecommunication studies from Waseda University in 2003 and 2007, respectively. He worked at Waseda University and Tokyo University of Science. After joining Tokyo University of Information Sciences as an Assistant Professor in 2011, he has been a Professor in the Department of Information Systems, Tokyo University of Information Sciences, since 2019. His research interests include network QoS control, network resource control and management, and network security. He is a member of the IEEE, IEICE and IPSJ.



Atsushi Waseda received the B.E. degree in communication engineering from the University of Electro-Communications in 2000. He received his M.S. and Ph.D. in information science from Japan Advanced Institute of Science and Technology (JAIST) in 2002, and 2007, respectively. He worked at the National Institute of Information and Communications Technology and KDDI research inc. After joining Tokyo University of Information Sciences as an Assistant Professor in 2019. His research interests include information security, quantum security and privacy protection. He is a member of the IEICE and IPSJ.



Moo Wan Kim received B.E., M.E. and Ph.D degree in electronic engineering from Osaka University, Osaka, Japan in 1974, 1977 and 1980, respectively. He joined Fujitsu Lab. in 1980 and had been engaged in research and development on multimedia communication systems, Intelligent Network, ATM switching system and operating system. In 1998 he joined Motorola Japan and had been engaged in research and development on CDMA2000 system. In 2000 he joined Lucent Japan and had been engaged in research and development on W-CDMA system, IMS and Parlay. In 2005 he joined Tokyo University of Information Sciences and had been engaged in research on Ubiquitous Network. In 2022 he joined TA Tech. as a lecture.