Regular Paper

Empowering Anti-malware Research in Japan by Sharing the MWS Datasets

Mitsuhiro Hatada[1,2,a),b)] Mitsuaki Akiyama[3,c)] Takahiro Matsuki[4,d)]
Takahiro Kasama[5,e)]

Received: December 1, 2014, Accepted: June 5, 2015

Abstract: Substantial research has been conducted to develop proactive and reactive countermeasures against malware threats. Gathering and analyzing data are widely accepted approaches for accelerating the research towards understanding malware threats. However, collecting useful data is not an easy task for individuals or new researchers owing to several technical barriers, such as conducting honeypot operations securely. The anti-Malware engineering WorkShop (MWS) was organized in 2008 to fill this gap; since then, we have shared datasets that are useful for accelerating the data-driven anti-malware research in Japan. This paper provides the definitive collection of the MWS Datasets that are a collection of different datasets for use in anti-malware research. We also report the effectiveness of the MWS Datasets from the viewpoint of published research papers and how to empower some of the papers by using the MWS Datasets. Furthermore, our discussion about issues of the MWS Datasets reveal the future directions for accelerating anti-malware research from the perspectives of dataset collection activity and dataset use activity.

Keywords: anti-Malware engineering WorkShop, MWS Datasets

1. Introduction

Substantial research has been conducted to develop proactive and reactive countermeasures against malware threats. Gathering and analyzing data are widely accepted approaches for accelerating the research towards understanding malware threats. Recent years have seen many efforts to take measures against evasive malware [1], [2] that avoid analysis and/or detection and to probe actively malicious servers [3], [4] such as those involved in command and control. These research achievements are based on the observation of massive real-world data.

However, collecting useful data is not an easy task for individuals or new researchers owing to technical barriers involving issues such as the types of honeypot they use, how to deploy their honeypot in the network environment, how many honeypots they operate, which tools they use to analyze the collected malware and so on. In addition, the results of analyzing collected data are helpful for use in additional research such as malware classification based on machine learning using dynamic analysis logs. To obtain accessible data, researchers must deploy an analysis environment and massively analyze the data as preparation. Not only technical obstacles, but also a risk of being exposed to and experiencing attacks is potentially involved.

In 2008, the anti-Malware engineering WorkShop (MWS) was organized in an effort to overcome these barriers. The main objective of the MWS is to accelerate and expand the data-driven anti-malware research in Japan by sharing useful datasets. In addition, research achievements presented at the MWS result in intensified efforts owing to the competition involved [5]. We have shared a summary of the MWS Datasets in Japanese [6], [7], [8], [9], [10], [11] covering the three attack phases of 1) probing, 2) infection, and 3) malware activities after infection, as shown in **Fig. 1**. And the competition [12] for anti-malware technique was started in 2009 as an application of the MWS Datasets. Throughout the seven years of MWS's history, the administrative and technical aspects of MWS 2009 were evaluated at the very beginning of the MWS in Ref. [13], and the experiences of sharing the datasets in the MWS community for a period of seven years were recently reported in Ref. [14]. Of course, there are

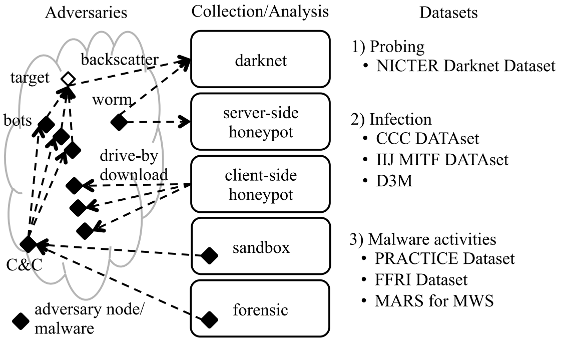


Fig. 1 Attack phases of malware applicable to the MWS Datasets.

[1] NTT Communications Corporation, Minato, Tokyo 108–8118, Japan
[2] Waseda University, Shinjuku, Tokyo 169–8555, Japan
[3] NTT Secure Platform Laboratories, Musashino, Tokyo 180–8585, Japan
[4] FFRI, Inc., Shibuya, Tokyo 150–0013, Japan
[5] National Institute of Information and Communications Technology, Koganei, Tokyo 184–8795, Japan
[a)] m.hatada@ntt.com
[b)] m.hatada@nsl.cs.waseda.ac.jp
[c)] akiyama.mitsuaki@lab.ntt.co.jp
[d)] matsuki@ffri.jp
[e)] kasama@nict.go.jp

many efforts [15], [16], [17], [18], [19], [20], [21], [22], [23] using the MWS Datasets, but the definitive collection of the details of each dataset and the practical effectiveness of sharing the MWS Datasets have never been comprehensively provided. This paper makes the following contributions:

- We provide the *definitive collection* of the MWS Datasets. The details of each dataset are described in a unified and comprehensive way so as to be referred to by many researchers. We believe that our paper is useful especially for a researcher who is about to start his/her research using the MWS Datasets because it is crucial for such a user to fully understand the data collection environment and techniques used to compile the dataset. Without such background information, he/she may fail to catch the limitations of the dataset, resulting in a wrong use of the datasets.
- We demonstrate the *effectiveness* of the MWS Datasets from the viewpoint of research activities that arise from the datasets. We first track the number of published papers presented at the MWS sessions. Next, we pick some research papers whose contributions have been empowered by the MWS Datasets. It is now easier than ever, not only to use the common dataset for their evaluation experiments but also to share their analysis techniques and results among researchers. These results definitively show that anti-malware research in Japan has been accelerated.
- In order to further dataset collection activity, we express the *significant concerns* based on our experience from four perspectives; dataset varieties, dataset volume, dataset continuity, and shared analysis environment. Our discussions open new opportunities, not only to accelerate the data-driven anti-malware research with researchers coming from a variety of fields, but also to collaborate with other communities to exchange useful datasets on global terms.

The remainder of this paper is structured as follows. Section 2 describes the details of each dataset included in the MWS Datasets. Section 3 shows the effectiveness of sharing the MWS Datasets by a record of usage trends of the MWS Datasets from the first usage in 2008 to the most recent in 2014 and research examples that each dataset contributed. Section 4 discusses the current challenges to anti-malware research and the required datasets. Section 5 introduces related work on research communities and datasets. Section 6 presents conclusions.

2. MWS Datasets

The MWS Datasets are a collection of different types of datasets that are designed for use in anti-malware research. We share the MWS Datasets within the MWS community, which consists of researchers in industry, academia, and research institutes in Japan. Many research achievements from the use of the MWS Datasets have been presented at the MWS session as a part of the Computer Security Symposium (CSS), which is one of the largest academic conference on computer security in Japan. By using a common dataset, researchers can start to work without performing initial data collection themselves and can learn practical techniques from others easily.

In expectation of sharing research achievements from diverse points of view at the MWS session, we assumed three type of researchers when we developed the first CCC DATAset in 2008; 1) highly-professional on malware analysis, 2) packet analysis expert such as intrusion detection, and 3) entry level researcher who is unfamiliar with malware analysis and/or packet analysis. In response to these types, we developed each dataset as in Sections 2.1.1, 2.1.2, and 2.1.3. We helped researchers to conduct continuous experiment and analyze long-term trend changes by providing past dataset after that.

Table 1 Answer for questions related to effectiveness of the MWS Datasets.

Question	Year	Yes	No	Neutral	No response
Q1	2008	11	2	1	1
	2009	19	1	5	0
	2010	17	1	6	0
Q2	2008	14	0	1	0
	2009	21	2	1	1
	2010	20	0	5	0

In order to improve the MWS Datasets based on requests from researchers, we had a questionnaire after each MWS session from 2008 to 2010. One of the free description type questionnaire was what dataset researcher wants. Examples of answer were following threat changes, increasing in volume, being analyzed by typical tool, being observed in different environment, being available in real time, being able to evaluate false positives, and the like. In addition, we hold a meeting before each MWS session every year to explain the MWS Datasets and expected research to researchers. In doing so, we discussed about the MWS Datasets face-to-face, received feedback, and asked researcher who might come around to our activity to provide a dataset. Another objective of a questionnaire was to measure the effectiveness of the MWS Datasets. We had two multiple-choice questions as follows; Q1) Did you come to be able to conduct new research by using the MWS Dataset? and Q2) Did you find a research challenge or research objective from paper presented at the MWS session? Total response was 15 in 2008, 25 in 2009, and 25 in 2010 and **Table 1** shows the result. The value of the MWS Datasets was clear from these figures. As a result through these activities, we have been motivated to develop and provide the MWS Datasets over the last seven years.

Table 2 shows the MWS Datasets catalogue by collection year, data format, and data size. Each dataset has interesting features useful for conducting advanced research. To assist researchers in using each dataset with a correct understanding, this section explains the details of each dataset from the viewpoints of collection environment, applied techniques, and expected research.

2.1 CCC DATAset

The CCC DATAset consists of a malware sample, honeypot packet trace, and malware collection log. *CCC* [24], [25], *Cyber Clean Center*, was a project that has analyzed characteristics of bots and botnets and provided information for the removal of bots from users' computers. The project was coordinated by the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry in Japan and ended in March 2011. This dataset was collected from server-side, high-interaction honeypots operated by *the CCC* in a distributed manner. Signifi-

Table 2 MWS Datasets Catalogue.

name	collection year	format	size
CCC DATAset (See Section 2.1)			
malware sample	2008~2013	tsv	18 K samples
packet traces (honeypot)	2008~2011	pcap	10 GB
malware collection log	2007~2011	csv	6.7 M records
MARS for MWS (See Section 2.2)	2008~2010	misc.	10 GB (compressed)
D3M (See Section 2.3)			
malware sample	2010~2014	tsv	73 samples
packet traces (honeypot)	2010~2014	pcap	219 MB
packet traces (dynamic analyis)	2012~2014	pcap	16 MB
IIJ MITF DATAset (See Section 2.4)			
attack communication log	2011~2012	csv	108.6 M records
malware collection log	2011~2012	csv	34.6 M records
PRACTICE Dataset (See Section 2.5)	2013	pcap	742 MB
FFRI Dataset (See Section 2.6)			
dynamic analysis log (oss)	2013~2014	json	5.6 K samples
dynamic analysis log (commercial)	2014	json	3 K samples
NICTER Darknet Dataset (See Section 2.7)	2011~2014	pcap	53 GB

cant features of these honeypots are the large scale, the use of multiple internet service providers (ISPs), and periodic reverting. Over a hundred honeypots accepted attacks and collected malware through multiple ISPs, creating a dataset that would be difficult for individual researchers to obtain. These honeypots were based on Windows 2000 and Windows XP SP1 virtual machines which periodically reverted at intervals of typically fifteen minutes to keep the environment safe but allowing malware to execute itself behind a firewall. Consequently, the packet trace contains malware activities occurring after the actual infection.

2.1.1 Malware Samples

A malware sample is used to perform research on malware analysis techniques. To select malware samples to be used for research, we set the three criteria; 1) Samples that have already been analyzed statically, 2) the samples should have intrinsic characteristics and functions, and 3) samples which are not detectable by major anti-virus software when they were collected. We applied these criteria year-by-year, resulting in 101 malware samples selected from 2008 to 2011. After 2012, the *CCC* liaison committee continued to operate honeypots on a reduced scale, even though the *CCC* project had ended. In accordance with this change, we revised third criteria; 3) that malware samples were detectable by major anti-virus software, which resulted in over 10 K malware samples for 2012 and 7 K malware samples for 2013.

To ensure the safe sharing of malware samples, the samples are specified as the lists of a hash digest. Actual malware binaries were made available appropriately within the MWS community.

2.1.2 Packet Traces

Packet traces of two honeypots based on Windows 2000 and XP SP1 for two successive days were provided in 2008 and 2009 to improve bot infection detection and analysis techniques. The collection period was extended from two days to one week in 2010 with both honeypots based on Windows XP SP1 due to Windows 2000 having reached its end-of-life. Further extension in 2011 involved the packet traces of two honeypots being gathered for two weeks twice a year, in August 2010 and in January 2011, aimed at detecting both short- and long-term trend changes. As we mentioned at the beginning of this subsection, packet traces contain malware activities occurring after infection during the reverting period of a honeypot. As a safeguard against unintended honeypot compromise, we restricted several ports accessible to the Internet and limited the traffic rate. Under these adjustments, malware activities after infection were very useful for identifying command and control communication, analyzing the downloading of further malware, and the like.

2.1.3 Malware Collection Logs

The malware collection log was offered for conducting research such as trend analysis techniques of bot and botnet activities from wider and long-term perspectives. In fact, the logs were collected in Japan by approximately a hundred honeypots under multiple ISPs for specifically three years and three months, from November 2007 to January 2011. Each record of the log consists of a time stamp, source IP address, source port number, destination IP address, destination port number, TCP or UDP, SHA1 hash digest of the collected malware binary, detection name by anti-virus software, and file path of the installed malware. The source or destination IP address representing a honeypot depending on the direction of communication was replaced with a honeypot ID such as honey016 to anonymize the location of the honeypot.

2.2 MARS for MWS

The *MARS*, *Malware/Minimal-attack Analysis Result Set* [26], is a set of dynamic and static analysis data for the malware samples of the CCC DATAset 2008, 2009, and 2010. One part of the metadata contains the hash digest, file name, file size, file type, detection name, and time stamp of the anti-virus software, as well as the version of the antivirus software with regard to a particular malware sample. The other is a reference list of each analysis result file with its analysis time stamp and tool information. These metadata are provided in XML format for researchers to use mechanically. Analysis results are classified on a dynamic and static basis.

Dynamic analysis is conducted on a non-virtualized physical server using a fake DNS server. The analysis results include the physical memory dump acquired after one minute of malware binary execution, packet traces for the five minutes following malware execution, and the query log of the fake DNS server. The results of static analysis consist of the strings of printable characters in the malware binary and several analysis results obtained by adapting the *Volatility Framework* [27] to the memory dump obtained in dynamic analysis, such as *dlllist, modules, splits, and sockets*. Since the MARS for MWS is a first-step analysis from both the dynamic and static perspectives, researchers can reduce the initial cost of malware analysis.

2.3 D3M

The Drive-by Download Data by *Marionette* (D3M) is a set of packet traces collected from the web-client, high-interaction honeypot system, *Marionette* [28], [29], which is based on Internet Explorer on Windows OS with several vulnerable plugins, such as Adobe Reader, Flash Player, WinZip, QuickTime and Java. This data focuses on drive-by download attacks caused by crawling malicious web sites according to threat transitions from exploiting OS services remotely. The datasets contain packet traces for the two periods; infection and after infection. The latter uses the dynamic malware analysis system, *BotnetWatcher* [30].

The packet traces of a web client honeypot are intended to identify the techniques that attackers are using. For example, a defaced web page contains a script to determine the version of a specific browser and its plugins and then redirects the client to a different web page to exploit the client effectively. The packet traces are collected after *Marionette* crawls the malicious URLs one time to filter out the benign pages. Since the data collection was conducted across different days in one year from 2010 to 2014, researchers can perform comparative analysis to determine the trend of the attack methods.

Malware is installed and executed on a host as a result of a drive-by download attack. However, *Marionette* does not allow malware execution after installation. The hash digests of collected malware binaries are listed in this data. Of course, real binaries can be obtained through the MWS community and researchers can try to analyze the malware itself, any correlation with the attack methods, and the like.

After malware collection by *Marionette*, each malware is analyzed for ten minutes in the *BotnetWatcher*, a dynamic analysis system connected to the Internet. Packet traces are captured during analysis; hence, they contain the malware's network activities, such as Internet connectivity checks, command and control communication, and the download of other malware. These are extremely useful for conducting extrusion detection research.

2.4 IIJ MITF DATAset

The IIJ MITF DATAset is collected by server-side, low-interaction honeypots based on the open source honeypot, *dionaea* [31], operated by *Internet Initiative Japan - Malware Investigation Task Force* (*IIJ MITF*) [32]. This dataset contains attack communication and malware collection logs from a hundred honeypots between July 1, 2011 and April 30, 2012 in order to discover the trends of bot and botnet. In addition, the collection period overlapped with the CCC DATAset, with the result that researchers can correlate attacks and analyze variations in the differences among the types of honeypot. Also, the CCC DATAset is based on honeypots distributed over multiple ISPs in Japan whereas the honeypots of *IIJ MITF* are deployed closely within one ISP.

The attack communication log contains necessary data such as time stamp, source IP address, source port number, honeypot ID, and destination port. The malware collection log is similar to the malware collection log of the CCC DATAset, except for the file type of downloaded malware and the exploited vulnerability information. As a kind of statistic, the concordance rate between the IP address that a honeypot was attacked by and IP address that a honeypot downloaded malware from is 99% which means that most of the malware attacks the host and then lets the host download malware from itself.

2.5 PRACTICE Dataset

The *PRACTICE* [33] project, *Proactive Response Against Cyber-attacks Through International Collaborative Exchange*, is funded by the Ministry of Internal Affairs and Communications and observes changes in the operation of the attackers for the defense response as one of their activities. The project's approach focuses on the long-term network activity of malware on a dynamic analysis system [30] connected to the Internet. For instance, an attacker updates the malware to improve functionality or deploys a redundant configuration of command and control servers.

The PRACTICE Dataset contains the packet traces obtained during long-term dynamic analysis of five malware samples (*Zbot*, *SpyEye*, etc.) and their metadata, containing a hash digest of the target malware, file name of packet traces, IP address assigned to the analysis host, analysis start and end times, and name detected by four anti-virus software packages. During an analysis period of up to one week at almost the same time in the middle of May 2013, we can observe peer-to-peer communication using high ports, downloading of files, repeated resolution of specific host names, and the like. For example, a specific piece of malware tried to join an IRC channel containing a specific user but it was blocked by server. The malware continued to attempt to rejoin periodically. In this case, no other activities were observed, so the amount of data for one week was 2.6 MB. Another example exhibited a large sized data collection of 494 MB, which contained massive peer-to-peer communication of *ZeroAccess* [34]. Analysis of malicious network activities, proposal of detection, and improvement of analysis environment are research examples expected using this dataset.

2.6 FFRI Dataset

The FFRI Dataset focuses on the internal activities that are caused to the host by malware and is provided in the JSON format that is generated by the dynamic analysis system. The *Cuckoo sandbox* [35] was used as an open source dynamic analysis tool in 2013 and 2014, and *FFR yarai analyzer Professional* [36], a commercial dynamic analysis system, was also used in 2014. The total number of analysis subjects was 2,644 in 2013 and 3,000 in 2014, solely in the Portable Executable (PE) format. Most of these were selected randomly from a massive collection of malware by crawling the web sites reflecting the trends of the malware at that time. In addition, there were a few malware related to well-known incidents such as a massive computer shutdown in South Korea on March 20, 2013 [37]. The execution time for *Cuckoo sandbox* was set to 90 seconds in 2013 and 120 seconds in 2014, respectively. For *FFR yarai analyzer Professional* [36], the execution time was set to 60 seconds.

Cuckoo sandbox generates an analysis report that consists of various information such as analysis time, analysis summary, file creation information, API calling, process tree, network activities, static information (hash digests, the section structure of PE file, import DLLs and the strings of printable characters in the malware binary), detection result by VirusTotal [38], and detection result by the default rule sets of the YARA signature [39]. The analysis summary, which is on the basis of file and registry access during malware execution, is useful to grasp an overview of malware behaviors. API call history is useful to analyze the details of malware activities.

FFR yarai analyzer Professional generates information logs that consist of histories of file access, registry access, network access, and API calls. It is noteworthy that this commercial tool has the capability of analyzing evasive malware [2] that can detect specific malware analysis environments and hide its malicious actions. Researchers can use the FFRI dataset to perform analysis such as classifying malware samples using the patterns of their activities, studying how the different dynamic analysis systems are affected by analysis environment, and etc.

2.7 NICTER Darknet Dataset

The NICTER Darknet Dataset is a set of packet traces collected from April 1, 2011 using the darknet monitoring system, *NICTER* [40]. The *NICTER* covers approximately 240 K unused IP addresses. Of the 240 K addresses, a subset of addresses with a /20 network prefix (4,096 IP addresses) is used for the NICTER Darknet Dataset. The packet traces contain scan packets to explore the reachable hosts by worms and researchers, backscatter packets caused by source IP address spoofing and the like. Thus, various suspicious activities can be observed by analyzing this dataset, which is also suitable for trend analysis as a result of containing more than three successive years of data. To anonymize the darknet sensors, the first and second octets of a destination IP address are replaced appropriately. A featured difference between this dataset and others is the ability for researchers to access past and real-time data using the *NONSTOP* [41], Platform as a Service (PaaS) environment, to analyze cyber security-related data safety from a remote site. In case of using original code to analyze data in the *NONSTOP*, researchers can transfer their code and execute them by themselves. On the other hand, downloading files from the *NONSTOP* to a local host is strictly restricted by multiple filters and the files are saved for audit to prevent the leak of cyber security-related data that is prohibited to bring out from the *NONSTOP*.

The number of packets collected per day by *NICTER* was almost 1.5 M. It has reached up to 4.5 M due to distributed reflection denial of service (DRDoS) attacks using DNS and NTP servers since 2013. The number of observed IP addresses was fewer than 10 K prior to the end of 2013, but it almost doubled in 2014 in response to the number of collected packets. Interesting research, such as determining predictors of a massive attack and correlation analysis with other datasets to deepen the suspicious activities observed can be conducted using this dataset.

3. Practical Effectiveness of Sharing MWS Datasets

We present the effectiveness of the MWS Datasets by tracking the record of published papers at yearly MWS sessions that used the datasets. In addition, we introduce research examples to explain how to empower their research by the use of the MWS Datasets. These results are concrete evidence that the MWS Datasets have helped accelerate anti-malware research in Japan.

3.1 Tracking the Record of MWS Datasets Use

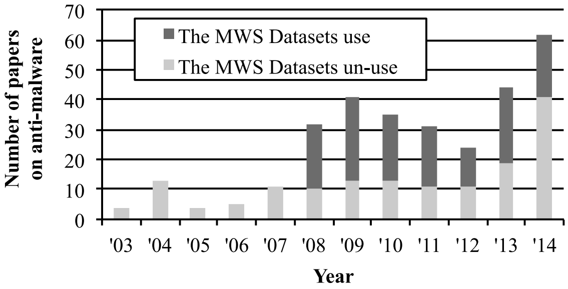


Fig. 2 Number of papers on anti-malware in CSS before and after sharing the MWS Datasets.

Table 3 Number of published papers using each dataset in MWS.

	’08	’09	’10	’11	’12	’13	’14
CCC DATAset	22	27	16	15	9	3	3
MARS for MWS	-	-	1	1	0	0	-
D3M	-	-	4	3	3	9	14
IIJ MITF DATAset	-	-	-	1	-	-	-
PRACTICE Dataset	-	-	-	-	-	3	1
FFRI Dataset	-	-	-	-	-	5	2
NICTER Darknet Dataset	-	-	-	-	-	6	2
Introduction of datasets	0	1	1	1	0	1	0
Total*	22	28	22	20	13	25	21
(Student paper)	8	15	10	9	9	10	10

-: Not available dataset that year included former dataset

*: Eliminated duplication of papers using multiple datasets

We measure the effectiveness of sharing the MWS Datasets by tracking the number of published papers on anti-malware at CSS from 2003 to 2014 before and after the first MWS as a part of CSS in 2008 (**Fig. 2**). The result shows that the total number of malware related papers has more than doubled since 2008 and that the MWS Datasets have significantly contributed to the expansion of the data-driven anti-malware research in Japan. **Table 3** shows the number of published papers using each dataset in past MWS. In addition to those in Table 3, there are some datasets that could not be provided owing to yearly contracts with dataset providers; these are shown as “-” in Table 3, as well as for the case that a certain dataset was yet to be introduced. There are some papers that used multiple datasets for their research, but the specified total number of papers eliminates duplications. The papers [6], [7], [8], [10] explaining the MWS Datasets were introduced to share the summary of datasets at the opening session of each workshop.

At the beginning of MWS, only the CCC DATAset was available to researchers, with papers using the CCC DATAset consisting mostly of papers appearing in MWS every year. In accordance with the threat transition to drive-by download attacks, the D3M could also be used, but the initial number of published papers was limited. However, the D3M has been completely accepted by researchers in the past two years, and research on drive-by download attacks has become a major trend in anti-malware research. The FFRI Dataset, NICTER Darknet Dataset, and PRACTICE Dataset were created in 2013. Although the number of papers using them decreased in 2014, these datasets contain a large amount of data compared to the CCC DATAset and the D3M; hence, new ways to use these datasets can be expected to be introduced in the near future. From another perspective, based on the number of student papers appearing each year, it is reasonable to say that the MWS Datasets are easier for non-professional people to use.

3.2 How to Empower Research

We have developed the MWS Datasets as a common dataset for researchers to start working without performing initial data collection themselves and learn practical techniques from others easily. To conduct continuous experiment and analyze long-term trend changes, we have provided various and past dataset, hence some research has been achieved to improve the reliability of proposed method and to share new findings from research. We present an abstract of a research example and explain how to empower such research by the use the MWS Datasets in each case. In summary, some research [16], [18], [20], [22], [23] has been conducted without initial effort to collect an evaluation dataset such as evaluating the proposed method properly, applying machine learning techniques, and improving the accuracy of the proposed method. Others [15], [17], [19], [21] have facilitated sharing the analysis techniques and results due to the effectiveness of using a common dataset such as providing a guide to develop the analysis environment, sharing the findings of analysis. This section is also helpful to give readers a better understanding of the research using each dataset.

CCC DATAset. *Alkanet* [15] can analyze malware using anti-debugging techniques to evade dynamic analysis tools. Tracing the system calls invoked by threads in real time can easily analyze a malware infection in another running process and help to understand malware behavior. In a few analysis cases, the behaviors of *SdBot*, *Palevo*, and *Polipos* in the CCC DATAset 2011 resolved such issues as the infection process, anti-debugging, and thread injection. The authors contributed by *sharing the technique and the detailed results* of analyzing the same dataset from a variety of real-world malware samples as one of the most basic concepts of dataset sharing.

In Ref. [16], packet traces of honeypots in the CCC DATAset 2009, 2010, and 2011 and packet traces of dynamic malware analysis in the D3M are used to evaluate the payload features for detecting the network activities of malware after infection. The authors *enhanced the reliability of the proposed method* by using not only a single dataset, but also multiple datasets of different years and types.

MARS for MWS. In Ref. [17], the difficulties in obtaining a dataset regarding the malware activities that would be observed in a real environment, and then the necessity of isolating a sandbox to avoid causing adverse side effects to the Internet from malware are explained. As *a good guide to improving the dynamic analysis of malware for safety*, the authors described the design and implementation of an isolated sandbox for analyzing malware using a mimetic Internet. Furthermore, the MARS for MWS was generated based on this sandbox using malware samples in the CCC Datasets 2008 and 2009. This research contributed by offering a useful example for generating datasets at an early stage of MWS history.

D3M. As an example of research on drive-by download attacks, a method based on applying an abstract syntax tree (AST) to characterize obfuscated malicious JavaScript code has been proposed [18]. The proposed method could classify similar obfuscated codes and detect similar codes by matching AST subtrees. The D3M is used to *evaluate the proposed method properly* without a barrier for collecting malicious web pages.

From another perspective, a paper [19] shows the results of analyzing HTTP communication data in the D3M, and *significant features of malicious redirections are found* effectively for malware detection. For example, the use of obfuscated JavaScript code with no referrer redirection can be used as the fingerprint of malicious redirection. Furthermore, the use of ephemeral port numbers without the normal port number 80 and the faked header field, "Content-Type," are not used on normal web pages.

IIJ MITF DATAset. This paper [20] focuses on identifying the command and control servers during the bot downloading phase. Although it is difficult to trace the source country/IP address of the botmaster, time zone correlation can be used as a tool to identify the time zone of the command and control servers. Using over 30 M data records and almost five hundred unique malware names from the IIJ MITF DATAset, the authors *found a strong correlation* between active bot downloads and the time zone of the command and control servers.

PRACTICE Dataset. In Ref. [21], features of malicious communication generated by malware are discussed in detail. The PRACTICE Dataset contains five files of packet traces during dynamic analysis of each malware. Hence, the paper reports analysis results such as longitudinal data analysis based on the communication protocol, host-based analysis of HTTP communication, DNS query analysis, periodicity analysis, and packet size analysis, as well as matching results with known IP address black lists for each subject file. *Some features enable distinguishing between normal traffic and malicious traffic*, but this effort remains confined to a survey; refer to the paper to understand this dataset.

FFRI Dataset. In Ref. [22], an unsupervised approach to extract *API call topics* from a large corpus of API calls is proposed. Through analysis of the API call logs collected from the thousands of malware samples of this dataset, the concept of "API call topics," representing a set of API calls intrinsic to a specific group of malware samples, is produced by conducting non-negative matrix factorization (NMF) clustering analysis. Detecting similar malware samples is demonstrated using extracted API call topics. Furthermore, relevant clustering results and the detected name from anti-virus software were discussed from an industrial perspective. The proposed method with wide use of the dataset contents is *a valuable case for applying machine learning techniques* to malware classification and behavioral detection of malware.

NICTER Darknet Dataset. Attention is currently focused on distributed DRDoS attacks, and amplifier probing for reflector attacks is analyzed using this dataset that is difficult to collect on their own in Ref. [23], with traffic captured from a campus network. This research aims to classify organizations or tools that perform amplifier probing and eliminate potential noise generated by the legitimate probing used for research purposes. The approach of extracting fingerprints to classify amplifier probing is based on case-control studies, and matching with the known IP address list of universities and research institutes enables the identification of legitimate probing. Results showed that this dataset causes the majority of legitimate probing for research purposes and the statistical differences between darknet and the campus network. Hence, these findings can *increase the accuracy of amplifier probing detection* and obtain information predictive of attacks by eliminating noise.

4. Discussion

We have developed and provided the MWS Datasets for last seven years as described in the beginning of Section 2. In this section, we explain current limitations regarding the MWS Datasets and discuss a path toward improving the MWS Datasets from our experiences in terms of dataset varieties, dataset volume, dataset continuity, and shared analysis environment. These discussions point out the future directions for accelerating anti-malware research and significance of dataset collection activity.

Dataset varieties. As we mentioned in Section 1, the MWS Datasets cover three attack phases from the perspective of adversaries. However, a recent attack vector can take many forms, depending on environmental changes to our lives. For example, smartphones and tablet PCs have been widely used not only for business but also for pleasure in recent years. We face the threat of Android malware, and now a relevant dataset has become available from another project [42]. In other instances, we must focus on the actions of cyber attackers under targeted attack, such as advanced persistent threats (APT). The research dataset *Behavior Observable System 2014* (*BOS_2014*) [43] was introduced and its observation environment explained in MWS 2014. The *BOS_2014* comprises two cases of a malware sample, packet traces, and a process log collected from a virtual company that is a malware execution environment. The malware that were attached to e-mails are provided as well as the observed attacker's activities on the host within the virtual company. On the other hand, to evaluate false positives and true negatives of the detection method, datasets of normal traffic or benign files are necessary for verifying usefulness. Researchers typically use benign data collected in their own environment, therefore, the comparison of proposed methods is not an easy task at present. In addition, it is crucially important to develop a required dataset by applying research results or collaborating with different groups.

Dataset scales and adequate sampling. When a researcher performs data-driven research, it is ideal to have as much data as possible because small and/or biased data could result in a wrong conclusion. In other words, having *generic* data is necessary to create *reproducible* research results. However, it is not feasible to collect the entire set of data available on the Internet. To fill this gap, the data providers can take two approaches: (1) increasing the volume of datasets as much as they can afford and (2) reducing the volume of datasets without sacrificing the essence of the data. In both cases, it is crucial that we can safely assume that the collected dataset represents the reality; e.g., it covers various malware families, it covers various types of attacks, and/or it is *not* biased to a specific set of cases, etc.

For the first approach (1), the MWS Datasets include more than 17 K hash values of malware samples, which is comparable with the 12 K malware samples provided by the MALICIA Project [44]. However, given the scale of recent malware analysis research such as Ref. [45], where 8 million malware samples are used, we may need to increase the volume of the MWS Datasets to be able to verify that the analysis results are not biased to a specific set of data. We note that increasing the dataset volume requires a lot of additional effort such as resource management or storage volume. We leave these items for our future work.

For the second approach (2), more research on performing appropriate data sampling would be necessary i.e., instead of trying to increase the volume in a blind way, we can focus our attention on the areas that likely reflect the essential features. For instance as we mentioned in Section 2, while the *NICTER* darknet covers 240 K IP addresses, the NICTER Darknet Dataset covers 4,096 IP addresses. If we could assume that the sampled 4 K IP addresses are a part of a /20 network prefix in the 240 K IP space, the reduced size of the dataset should suffice for some objectives. Also, we may want to extend the coverage of the dataset, eliminating a huge number of potential duplications. Thus, the collection of a large amount of data together with the adequate sampling technologies is left for future research topics. In case of malware sample, if we can achieve effective sampling that reflected major features of malware, we can identify minor features or new features to put a high priority on analysis. As one of the future research items, it is also promising that researchers collaboratively collect data using the same tools and share the data with each other.

Dataset continuity. We must not lose sight of another perspective; the continuous accessibility of a dataset to researchers. Once a dataset is no longer provided, researchers are unable to continue to improve their proposed methods, such as anomaly detection or attack trend analysis. Nonetheless, there is cost in maintaining the observation environment, collecting a dataset, and preparing to provide the dataset. Obviously, a dataset is crucial for researchers to conduct advanced research, but sharing a dataset is not a simple task. Providing a dataset also presents the advanced technology of the dataset provider. However, providing a dataset continuously allows one to receive good feedback from researchers using the dataset. It can also serve as a trigger for promoting new research. From a long-term perspective, both uni- and bidirectionality can balance dataset providers and users in industry, academia, and research institutes.

Shared analysis environment. For further discussion, we mention the analysis environment for using the datasets. As we mentioned in Section 2.7, the NICTER Darknet Dataset is provided in a remote access environment to allow the safe access of darknet traffic. If commonly used tools are available in the shared environment in this way, then the barriers to reproducing existing approaches for comparison, when researchers evaluate their proposals, can be lowered. In addition, with respect to the dataset volume, the control of huge data might increase the cost and decrease the accessibility of a dataset. Thus, a shared environment of dataset and analysis tools is one idea for accelerating data-driven anti-malware research in the future.

5. Related Work

We review some examples of shared datasets in the research community in parallel with the attack phases noted in Fig. 1.

In phase 1) probing, there are various types of datasets. The CAIDA [46] collects several different types of data, including

Table 4 Comparison with typical dataset.

Dataset	Variety	Volume	Continuity
MWS	Yes	Yes	Yes
CAIDA [46]	No	Yes	Yes
PREDICT [47]	No	Yes	Yes
DARPA [49]	No	Yes	No
MALICIA [44]	No	-	No

backscatter, distributed denial of service (DDoS) scans, and worms, as well as traffic and topology. The PREDICT [47] shares 430 datasets in 13 categories contributed by nine data providers, such as Blackhole Address Space Data, IDS and Firewall Data, and SSL Certificate Data. Researchers in the USA and other selected countries are approved for creating accounts and accessing the repository. Since 2011, WOMBAT [48] has organized open workshops known as BADGERS workshops. This project aims to gather security related raw data, enrich the data using analytical techniques, and provide root cause analysis to project members.

In phase 2) infection, the historic DARPA Intrusion Detection Data Sets [49] from 1998 to 2000 were released training data and testing data for IDS evaluation with packet traces, audit data, and file system dumps. The hpfriends social data-sharing platform [40] of the Honeynet Project shares the distributed honeypots operated by each contributor. The Contagio [50] shares malware samples relating to APT attacks and their packet traces captured during dynamic analysis. The Open Malware [51] stores over five million samples as of November 11, 2014 collected by honeypots and user submissions. The MALICIA Project [44] provides 11,688 labeled malware samples collected over a period of 11 months and lists 42 institutions releasing this dataset on the web. The Android Malware Genome Project [42] shares over 1,200 Android malware samples. As of July 31, 2014, this project has been released to 421 universities, research labs, and companies listed on the web.

Datasets corresponding to phase 3) of malware activities are available from free services. The Anubis [52] is a free analysis service operated by a security research group. Users can submit a subject for analysis using Windows executables and Android APK files, and obtain an analysis report. The Malwr [53] is another free analysis service and community launched in 2011. All analysis subjects of the Malwr have totalled more than 211,000 as of November 11, 2014.

As there are many others, we cite a few more examples. The MACCDC [54] and the CDX [55] were collected during an attack and defense challenge by a Red Team versus a Blue Team, and contain packet traces and related logs. The CFReDS [56] is a set of forensic images for search data and file carving. Using synthetic data for machine learning-based cyber security experiments is discussed in [57].

Table 4 summarizes typical dataset [44], [46], [47], [49] introduced as above. *Variety* column denotes if the dataset covers three attack phases as we mentioned in Section 1. Only the MWS Datasets cover the three attack phases of 1) probing, 2) infection, and 3) malware activities after infection, as shown in Fig. 1 *Volume* column denotes if the dataset can be used for experimentation. Many publications using the dataset are listed in each website or in the result of the web search. The most recent MALICIA dataset has been provided since 2013 so it has potential for growth in the near future. *Continuity* column denotes the continuous accessibility and update of the dataset. Both the DARPA Intrusion Detection Data Sets and the MALICIA dataset are currently available but they are not renewed for the time being at least. All of these datasets can together fill a data gap with the MWS Datasets, however, the MWS Datasets have an advantage for anti-malware research from the perspectives of dataset variety and continuity.

6. Conclusion

In this paper, we introduced the MWS Datasets and shared useful datasets for anti-malware research covering the three attack phases of probing, infection, and malware activities after infection. To demonstrate the effectiveness of the MWS Datasets, we tracked the number of published papers on anti-malware at yearly CSS before and after sharing the MWS Datasets. The total number of malware-related papers more than doubled as a result, and therefore we can judge that we significantly contributed to accelerate anti-malware research in Japan. Also we presented some research cases to explain how to empower malware research by using the MWS Datasets. Lowering the barrier of the initial effort to collect a dataset and sharing analysis techniques and results were essential factors to improve research activities. By doing this, this paper provides the definitive collection of the MWS Datasets for overseas researchers to refer to in using the datasets. We also discussed the varieties of dataset and their continuous accessibility for the purposes of improving the datasets. Not only the datasets, but also the analysis environment and dataset sizes were discussed to accelerate anti-malware research. Even if these datasets are easy to obtain, it is also necessary to improve the ability of developing datasets. Thus, the sustainable framework of collecting, sharing, and utilizing the datasets with the contribution of each other is indispensable to anti-malware research. Of course, we continue efforts to discuss and develop new datasets depending on the trend of attacks. Hereafter, we hope that researchers will develop new datasets to share by applying their own techniques and/or environments.

Acknowledgments We thank National Institute of Information and Communications Technology for providing the MARS for MWS, Internet Initiative Japan Inc. for providing the IIJ MITF DATAset, and all the members of the MWS community.

References

[1] Kapravelos, A., Shoshitaishvili, Y., Cova, M., Kruegel, C. and Vigna, G.: Revolver: An Automated Approach to the Detection of Evasive Web-based Malware, *22nd USENIX Security Symposium* (*USENIX Security 13*), pp.637–652 (2013).

[2] Kirat, D., Vigna, G. and Kruegel, C.: BareCloud: Bare-metal Analysis-based Evasive Malware Detection, *23rd USENIX Security Symposium* (*USENIX Security 14*), pp.287–301 (2014).

[3] Nappa, A., Xu, Z., Rafique, M.Z., Caballero, J. and Gu, G.: CyberProbe: Towards Internet-Scale Active Detection of Malicious Servers, *Proc. 21st Annual Network and Distributed System Security Symposium* (*NDSS'14*), pp.1–15 (2014).

[4] Xu, Z., Nappa, A., Baykov, R., Yang, G., Caballero, J. and Gu, G.: AutoProbe: Towards Automatic Active Malicious Server Probing Using Dynamic Binary Analysis, *Proc. 21st ACM Conference on Computer and Communications Security* (*CCS'14*), pp.1–12 (2014).

[5] Terada, M.: Anti Malware engineering WorkShop (MWS) Contributions and challenges to the educational community, IPSJ SIG Techni-

cal Reports 2011-CE-112 10 (2011), (in Japanese).

[6] Hatada, M., Nakatsuru, Y., Terada, M. and Shinoda, Y.: Datasets for anti-malware research and research achievements shared at the workshop, *Proc. Computer Security Symposium 2009* (*CSS2009*), Vol.2009, pp.1–8 (2009), (in Japanese).

[7] Hatada, M., Nakatsuru, Y., Akiyama, M. and Miwa, S.: Datasets for Anti-Malware Research —MWS 2010 Datasets, *Proc. Computer Security Symposium 2010* (*CSS2010*), Vol.2010, pp.1–5 (2010), (in Japanese).

[8] Hatada, M., Nakatsuru, Y. and Akiyama, M.: Datasets for Anti-Malware Research —MWS 2011 Datasets, *Proc. Computer Security Symposium 2011* (*CSS2011*), Vol.2011, No.3, pp.1–5 (2011), (in Japanese).

[9] MWS 2012, available from ⟨http://www.iwsec.org/mws/2012/about.html⟩ (accessed 2014-11-28), (in Japanese).

[10] Kamizono, M., Hatada, M., Terada, M., Akiyama, M., Kasama, T. and Murakami, J.: Datasets for Anti-Malware Research —MWS Datasets 2013, *Proc. Computer Security Symposium 2013* (*CSS2013*), Vol.2013, No.4, pp.1–8 (2013), (in Japanese).

[11] Akiyama, M., Kamizono, M., Matsuki, T. and Hatada, M.: Datasets for Anti-Malware Research —MWS Datasets 2014, IPSJ SIG Technical Reports 2014-CSEC-66 19 (2014), (in Japanese).

[12] Takemori, K. and Hosoi, T.: MWS Cup 2009 (Special Feature: Malware), *IPSJ Magazine Joho Shori*, Vol.51, No.3, pp.296–299 (2010), (in Japanese).

[13] Hosoi, T. and Matsuura, K.: Evaluation of the Common Dataset Used in Anti-Malware Engineering Workshop 2009, *Proc. 13th International Symposium, RAID 2010*, Vol.6307, pp.496–497 (2010).

[14] Hatada, M., Terada, M. and Mori, T.: POSTER: Seven Years in MWS: Experiences of Sharing Datasets with Anti-malware Research Community in Japan, *Proc. 2014 ACM SIGSAC Conference on Computer and Communications Security* (*CCS'14*), pp.1433–1435 (2014).

[15] Otsuki, Y., Takimoto, E., Kashiyama, T., Saito, S., Cooper, E. and Mouri, K.: Tracing Malicious Injected Threads Using Alkanet Malware Analyzer, *IAENG Trans. Engineering Technologies*, Lecture Notes in Electrical Engineering, Vol.247, pp.283–299 (2014).

[16] Otsuki, Y., Masatsugu, I., Soichi, K., Mitsuhiro, H. and Hiroshi, Y.: Evaluating payload features for malware infection detection, *Journal of Information Processing*, Vol.22, No.2, pp.376–387 (2014).

[17] Miwa, S., Kadobayashi, Y. and Shinoda, Y.: Generate Activity Dataset using MAT; Minimal-attack/Malware Analysis Testbed, *Proc. Computer Security Symposium 2009* (*CSS2009*), Vol.2009, pp.1–6 (2009).

[18] Blanc, G., Miyamoto, D., Akiyama, M. and Kadobayashi, Y.: Characterizing Obfuscated JavaScript Using Abstract Syntax Trees: Experimenting with Malicious Scripts, *Proc. 26th Advanced Information Networking and Applications Workshops* (*WAINA*), pp.344–351 (2012).

[19] Takata, Y., Goto, S. and Mori, T.: Analysis of Redirection Caused by Web-based Malware, *Proc. 32nd Meeting of the Asia-Pacifc Advanced Network*, pp.53–62 (2011).

[20] Sisaat, K., Kikuchi, H., Kittitornkun, S., Yukonhiatou, C., Terada, M. and Ishii, H.: Time Zone Analysis on IIJ Network Traffic for Malicious Botnet Activities, Technical Report of IEICE 135 (2013).

[21] Tanaka, Y., Hatada, M. and Inazumi, T.: A Study of Characteristic of Malignant Communication as Seen from the Packet Capture Data, *Proc. Computer Security Symposium 2013* (*CSS2013*), Vol.2013, No.4, pp.118–124 (2013).

[22] Fujino, A., Murakami, J. and Mori, T.: Discovering Similar Malware Samples Using API Call Topics, *Proc. 2015 IEEE Consumer Communications and Networking Conference* (*CCNC 2015*) (2015), (to appear).

[23] Haga, Y., Saso, A., Mori, T. and Goto, S.: Analysis of Amplifier Probing for Reflector Attacks, *Proc. Computer Security Symposium 2014*, Vol.2014, No.2, pp.300–307 (2014).

[24] Cyber Clean Center, available from ⟨https://www.telecom-isac.jp/ccc/en_index.html⟩ (accessed 2014-11-28).

[25] Arimura, K.: The Cyber Clean Center Project, Practical Cooperation of Major Japanese ISPs Anti-botnet Activity (Special Feature: Malware), *IPSJ Magazine Joho Shori*, Vol.51, No.3, pp.275–283 (2010), (in Japanese).

[26] Miwa, S., Miyachi, T., Eto, M., Yoshizumi, M. and Shinoda, Y.: Design and Implementation of an Isolated Sandbox with Mimetic Internet used to Analyze Malwares, *Proc. DETER Community Workshop on Cyber Security Experimentation and Test*, pp.1–9 (2007).

[27] The Volatility Framework, available from ⟨https://code.google.com/p/volatility/⟩ (accessed 2014-11-28).

[28] Akiyama, M., Aoki, K., Kawakoya, Y., Iwamura, M. and Itoh, M.: Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks, *IEICE Trans. Commun.*, Vol.E93-B, No.5, pp.1131–1139 (2010).

[29] Akiyama, M., Yagi, T., Kadobayashi, Y. and Hariu, T.: Client Honeypot Multiplication with High Performance and Precise Detection, *IEICE Trans. Info. Sys.*, Vol.E98-D, No.4, pp.775–787 (2015).

[30] Aoki, K., Yagi, T., Iwamura, M. and Itoh, M.: Controlling malware HTTP communications in dynamic analysis system using search engine, *Proc. 3rd International Workshop on Cyberspace Safety and Security*, pp.1–6 (2011).

[31] dionaea, available from ⟨http://dionaea.carnivore.it/⟩ (accessed 2014-11-28).

[32] IIJ MITF, available from ⟨https://sect.iij.ad.jp/en/mitf.html⟩ (accessed 2014-11-28).

[33] PRACTICE: Proactive Response Against Cyber-attacks Through International Collaborative Exchange, available from ⟨http://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/Releases/Telecommunications/130307_02.html⟩ (accessed 2014-11-28).

[34] The ZeroAccess Botnet: Mining and Fraud for Massive Financial Gain, available from ⟨http://www.sophos.com/en-us/why-sophos/our-people/technical-papers/zeroaccess-botnet.aspx⟩ (accessed 2014-11-28).

[35] Cuckoo sandbox, available from ⟨http://www.cuckoosandbox.org/⟩ (accessed 2014-11-28).

[36] FFR yarai analyzer Professional, available from ⟨http://www.ffri.jp/products/yarai_analyzer_pro/⟩ (accessed 2014-11-28), (in Japanese).

[37] Dissecting Operation Troy: Cyberespionage in South Korea, available from ⟨http://www.mcafee.com/sg/resources/white-papers/wp-dissecting-operation-troy.pdf⟩ (accessed 2014-11-28).

[38] VirusTotal, available from ⟨https://www.virustotal.com/⟩ (accessed 2014-11-28).

[39] YARA, available from ⟨http://plusvic.github.io/yara/⟩ (accessed 2014-11-28).

[40] Nakao, K., Inoue, D., Eto, M. and Yoshioka, K.: Practical Correlation Analysis between Scan and Malware Profiles against Zero-Day Attacks Based on Darknet Monitoring, *IEICE Trans. Inf. Syst.*, Vol.E92-D, No.5, pp.787–798 (2009).

[41] Takehisa, T., Inoue, D., Eto, M., Yoshioka, K., Kasama, T., Nakazato, J. and Nakao, K.: NONSTOP: Secure Remote Analysis Platform for Cybersecurity Information, Technical Report 95 (2013), (in Japanese).

[42] Zhou, Y. and Jiang, X.: Dissecting Android Malware: Characterization and Evolution, *Proc. 33rd IEEE Symposium on S&P*, pp.95–109 (2012).

[43] Terada, M., Aoki, S., Kusumi, J., Shigemoto, T. and Hagihara, K.: Feasibility Study of Research Data Set "Behavior Observable System 2014", *Proc. Computer Security Symposium 2014*, Vol.2014, No.2, pp.1121–1125 (2014).

[44] Nappa, A., Rafique, M.Z. and Caballero, J.: Driving in the Cloud: An Analysis of Drive-by Download Operations and Abuse Reporting, *Proc. 10th Conference on DIMVA*, pp.1–20 (2013).

[45] Ye, Y., Li, T., Huang, K., Jiang, Q. and Chen, Y.: Hierarchical associative classifier (HAC) for malware detection from the large and imbalanced gray list, *Journal of Intelligent Information Systems*, Vol.35, No.1, pp.1–20 (2010).

[46] CAIDA Data, available from ⟨http://www.caida.org/data/overview/⟩ (accessed 2014-11-28).

[47] PREDICT, the Protected Repository for the Defense of Infrastructure Against Cyber Threats, available from ⟨https://www.predict.org/⟩ (accessed 2014-11-28).

[48] WOMBAT project: Worldwide Observatory of Malicious Behaviors and Attack Threats, available from ⟨http://www.wombat-project.eu/⟩ (accessed 2014-11-28).

[49] DARPA Intrusion Detection Data Sets, available from ⟨http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/⟩ (accessed 2014-11-28).

[50] Contagio, available from ⟨http://contagiodump.blogspot.jp/⟩ (accessed 2014-11-28).

[51] Open Malware, available from ⟨http://www.offensivecomputing.net/⟩ (accessed 2014-11-28).

[52] Anubis, available from ⟨https://anubis.iseclab.org/⟩ (accessed 2014-11-28).

[53] Malwr, available from ⟨https://malwr.com/⟩ (accessed 2014-11-28).

[54] The U.S. National CyberWatch Mid-Atlantic Collegiate Cyber Defense Competition (MACCDC), available from ⟨http://www.netresec.com/?page=MACCDC⟩ (accessed 2014-11-28).

[55] Sangster, B., O'Connor, T., Cook, T., Fanelli, R., Dean, E., Adams, J., Morrell, C. and Conti, G.: Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets, *USENIX Security's Workshop on Cyber Security Experimentation and Test* (*CSET*), p.9 (2009).

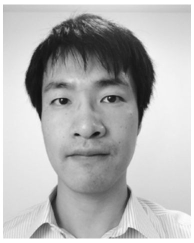
[56] Computer Forensic Reference Data Sets (CFReDS), available from ⟨http://www.cfreds.nist.gov/⟩ (accessed 2014-11-28).

[57] Abt, S. and Baier, H.: A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments, *Proc. 2014 Workshop on Artificial Intelligent and Security Workshop*

(*AISec'14*), pp.37–45 (2014).

Mitsuhiro Hatada was born in 1978. He is currently a Ph.D. student with particular interest in anti-malware. He received his B.E. and M.E. degrees in Computer Science and engineering from Waseda University in 2001 and 2003, respectively. He joined NTT Communications Corporation in 2003 and has been engaged in the R&D of network security and anti-malware. He is a member of IPSJ and IEICE.

Mitsuaki Akiyama received his M.E. and Ph.D. degrees in Information Science from Nara Institute of Science and Technology, Japan in 2007 and 2013, respectively. Since joining Nippon Telegraph and Telephone Corporation NTT in 2007, he has been engaged in research and development of network security, especially honeypot and malware analysis. He is now with the Network Security Project of NTT Secure Platform Laboratories.



Takahiro Matsuki received his B.E. in Engineering from Okayama University, Japan, in 2005. He received his Ph.D. in Engineering from Chuo University, Japan, in 2011. Since joining LAC Co., Ltd. in 2005, he has been working on information security and malware research. He is a CISSP since 2007. He has been contributing to MWS since 2008. He joined FFRI, Inc. in 2011, and he is currently director of Basic Research Laboratories.

Takahiro Kasama received his B.E., M.E. and Ph.D. degrees in Computer Engineering from Yokohama National University in 2009 and 2011, 2014, respectively. He is currently a researcher at the National Institute of Information and Communications Technology, Japan. His research interest covers a wide area of network security including malware analysis and network monitoring.