Poster: Seven Years in MWS: Experiences of Sharing Datasets with Anti-malware Research Community in Japan

Mitsuhiro Hatada[*‡]
[*]NTT Communications Corporation
Gran Park Tower 16F, 3-4-1 Shibaura Minato
Tokyo, Japan 108-8118
m.hatada@ntt.com

Masato Terada[†]
[†]Hitachi Incident Response Team
1-1-2 Kashimada, Saiwai, Kawasaki,
Kanagawa, Japan 212-8567
masato.terada.rd@hitachi.com

Tatsuya Mori[‡]
[‡]Waseda University
3-4-1 Okubo Shinjuku
Tokyo, Japan 169-8555
{m.hatada, mori}@nsl.cs.waseda.ac.jp

ABSTRACT

In 2008, the anti-Malware engineering WorkShop (MWS) was organized in Japan. The main objective of MWS is to accelerate and expand the activities of anti-malware research. To this end, MWS aims to attract new researchers and stimulate new research by lowering the technical obstacles associated with collecting the datasets that are crucial for addressing recent cyber threats. Moreover, MWS hosts intimate research workshops where researchers can freely discuss their results obtained using MWS and other datasets. This paper presents a quantitative accounting of the effectiveness of the MWS community by tracking the number of papers and new researchers that have arisen from the use of our datasets. In addition, we share the lessons learned from our experiences over the past seven years of sharing datasets with the community.

Categories and Subject Descriptors

K.6.5 [**MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS**]: Security and Protection—*Invasive software (e.g., viruses, worms, Trojan horses)*

General Terms

Security

Keywords

MWS, malware, dataset, research community

1. INTRODUCTION

In the field of anti-malware research, collecting and analyzing data is a widely established approach towards understanding this rapidly evolving target. To accelerate this highly data-driven research, it would be the most effective to stimulate new research and to attract new researchers from various disciplines, e.g., cyber security, networking, machine learning, and bioinformatics.

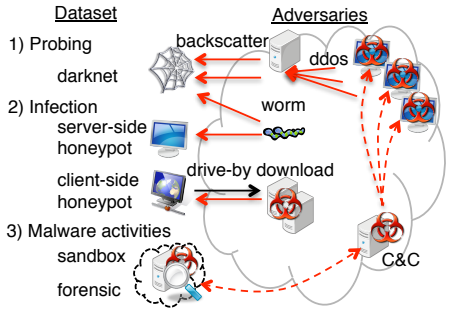


Figure 1: Attack phases of malware applicable to the MWS Datasets.

However, collecting useful data for anti-malware research is not an easy task for individual researchers because of several technical obstacles. For instance, operating a honeypot is a commonly used technique to collect malware. Although there are publicly available honeypot software packages, installing, configuring, and securely operating a honeypot generally requires considerable effort and experience.

To fill this gap, the anti-Malware engineering WorkShop (MWS) [1] was organized in 2008. The objective of MWS is to accelerate and expand the activities of anti-malware research by sharing community datasets among researchers. In addition, MWS has hosted intimate research workshops where researchers can discuss their research results obtained using MWS and other datasets. Moreover, to encourage student participation in the community, MWS also hosts competitions (the MWS Cup), which employ the MWS Datasets.

To facilitate the goals of stimulating research and attracting new researchers, the MWS Datasets have been developed with the following noteworthy features. First, the datasets are applicable to several attack phases: 1) probing, 2) infection, and 3) malware activities after infection, as illustrated in Fig. 1. Second, some of the datasets assist researchers in performing the long-term analysis. One of the datasets was collected from 2008 to 2013, and provides communication logs from a server-side, high-interaction honeypot. In addition, in response to attack vector transition, a drive-by download dataset has been provided since 2010. Finally, datasets have been developed to facilitate the correlation of various datasets collected by different research institutes and industries. For example, forensic data regarding phase

CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.
ACM 978-1-4503-2957-6/14/11.
http://dx.doi.org/10.1145/2660267.2662357.

3) malware activities is produced by analyzing malware samples collected during phase 2) infection.

The main contributions of this paper are as follows.

- We quantify the effectiveness of community data sharing by tracking the number of papers and new researchers that have arisen from the use of our datasets.
- We share the lessons learned from our experience over the past seven years of sharing datasets with the research community.

The remainder of this paper is organized as follows. Section 2 provides a brief summary of our datasets. In Section 3, we quantify the effectiveness of the MWS community by tracking the number of papers and new researchers that have arisen from the use of our datasets and discuss the lessons learned from our experiences. Section 4 discusses the related efforts to dataset sharing and Section 5 concludes our work.

2. MWS DATASETS

As shown in Fig. 1, the MWS Datasets cover three attack phases, i.e., probing, infection, and malware activities. Table 1 summarizes the datasets shared in the MWS community and their relationships. A brief overview of each dataset is provided below:

1) Probing.

The **NICTER Darknet Dataset** is a set of packet traces collected using the darknet monitoring system, *NICTER* [12]. Researchers can access realtime datasets using the Platform as a Service (PaaS) environment. The set of darknets covers approximately 210 K unused IP addresses.

2) Infection.

The **CCC DATAset** contains the data collected from server-side, high-interaction honeypots that are operated by the *Cyber Clean Center* [3] in a distributed manner. These datasets contain the list of hash digests for collected malware samples, packet traces collected on the honeypots, and the logs of malware collection.

The **IIJ MITF Dataset** is a set of logs collected from server-side, low-interaction honeypots operated by *MITF* [5]. As shown in Table 1 (a), this dataset can be directly correlated with the CCC DATAset because the data collection period and the format of logs are common among the two datasets.

The **D3M** is a set of packet traces collected from the web-client, high-interaction honeypot system, *Marionette* [9]. This data focuses on the drive-by download attacks of crawling malicious web sites. The datasets contain packet traces for two periods: at the time of infection and after the infection. The latter employs the dynamic malware analysis system, *Botnet Watcher* [10].

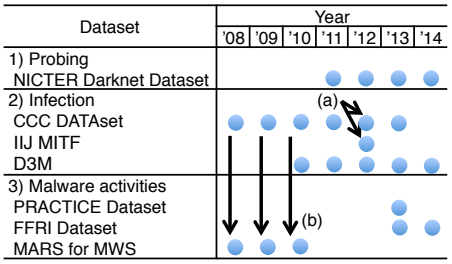
3) Malware activities.

The **PRACTICE Dataset** is a collection of long-term packet traces collected from the dynamic malware analysis system operated by the *PRACTICE* project [6]. The longest analysis period is approximately one week.

The **FFRI Dataset** is a set of logs collected from the dynamic malware analysis systems *Cuckoo sandbox* [2] and *yarai analyzer Professional* [4]. The analyzed malware samples are randomly chosen from large-scale malware archives collected from various sources.

The **MARS for MWS** is a set of malware dynamic analysis data collected from not-virtualized physical servers using a fake DNS server [11]. The dataset includes the memory dump and its forensic data. The malware samples analyzed in the MARS datasets were collected from the CCC DATAset (Table 1 (b)).

Table 1: Available datasets for each year.



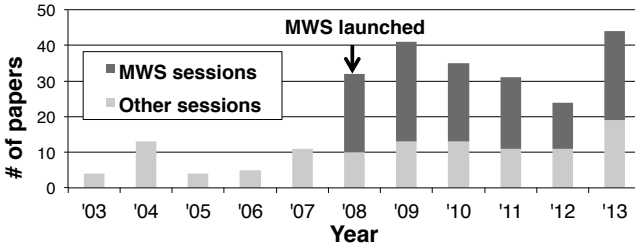


Figure 2: Number of published papers related to malware in CSS.

3. SEVEN YEARS OF EXPERIENCES

MWS has been held as a part of Computer Security Symposium (CSS), which is the largest domestic security research conference in Japan. Figure 2 presents the number of papers reflecting malware-related topics presented at CSS. The launch of MWS has significantly contributed to the increase in the number of anti-malware research papers. Interestingly, not only the number of papers presented at the MWS sessions but also the number of papers presented at other sessions has increased.

Table 2 illustrates the growth of the MWS community. The number of research groups in our community tripled from 2008 to 2014. Among these research groups, roughly 30 groups constantly made yearly contracts with the MWS organizing committee for the use of the datasets. We also counted the number of new research groups. The new research groups are those that have not worked in malware-related research in the past and their first paper on malware-related research was presented at MWS. From the results, we may conclude that MWS has successfully expanded the activities of anti-malware research over the past seven years.

Finally, Table 3 lists the outcomes of MWS with respect to the number of published papers that utilized MWS Datasets. Note that the numbers given for 2014 are as of July 12, 2014. In the past five years, the total number of publications has reached 30. It is also indicative of the effectiveness of the MWS community in accelerating anti-malware research.

We summarize our key successes obtained over the past seven years in the MWS community as follows.

- **Data**: among the datasets provided, packet traces have attracted the most newcomers. These datasets are suited for performing various analyses such as machine learning. In addition, the synchronization of the formats and collection periods of different datasets facilitates the identi-

Table 2: MWS community growth.

	'08	'09	'10	'11	'12	'13	'14
# of groups	28	48	54	59	71	83	86
# of groups w/contraction	25	27	33	26	30	38	31
# of new groups	2	5	2	2	3	3	-

Table 3: Number of published papers that have used MWS Datasets.

	'10	'11	'12	'13	'14	total
Journal (en)	0	2	0	4	1	7
Journal (ja)	2	1	2	3	1	9
Conference Proc.	4	3	5	2	0	14
subtotals	6	6	7	9	2	30

fication of common and separate trends of attack. Moreover, the types of datasets have been flexibly updated to remain abreast of threat transitions in the wild.

- **Lowering obstacles** : the MWS community has attempted to lower the various barriers to new research as much as possible. First, as previously mentioned, we lowered the technical obstacles of data collection by sharing datasets. Second, because our intent was to make the datasets available to any researcher wishing to conduct anti-malware research using datasets, we simplified the procedure for accessing these datasets as much as possible. We believe that these measures were effective for lowering the obstacles to new and enterprising research. Finally, to avoid the neglect of students who are less capable with the English language, we provided the descriptions of the datasets in Japanese. Thus, we lowered the barriers associated with language.

4. RELATED WORK

Among several shared datasets in the research community, we review some examples that run parallel to our own. The MALICIA Project [13] provides 11,688 labeled malware samples collected over a period of 11 months. The Android Malware Genome Project [14] shares over 1,200 Android malware samples. As of March 13, 2014, this project has been released to 370 universities, research labs, and companies. However, these two activities may be better referred to as research data repositories rather than research communities such as that of MWS. The closest activities to our own are PREDICT [7] in the USA and the WOMBAT project [8] in the EU. PREDICT shares 430 datasets consisting of 13 categories contributed by 9 data providers. Researchers in the USA and selected countries are approved to create accounts and to access the repository. WOMBAT has organized open workshops, known as BADGERS workshops, since 2011. This project aims to gather security related raw data, enrich the data by analytical techniques, and provide root cause analysis to project member. Unlike these two activities, MWS not only conducts workshops but also conducts competitions.

5. CONCLUSION

This paper has sought to describe our experiences of the past seven years with sharing datasets in the research community. The quantitative analysis regarding the growth of the research community and research outcomes demonstrated that MWS was able to accelerate and expand the activities of anti-malware research. For instance, 17 new research groups have arisen from the community and 30 research papers using MWS Datasets have been published so far. The experiences of the past seven years have revealed the effectiveness associated with lowering various barriers to entry in this field, i.e., facilitated data collection, simple procedure for accessing the datasets, and the reduction of language barriers. We believe that our experiences can assist other research communities that have a similar vision and comparable objectives. We are now planning to expand our activities to the global research community in response to several requests for accessing the MWS Datasets from researchers in other countries.

Acknowledgements

We thank all the members of the MWS community.

6. REFERENCES

[1] anti Malware engineering WorkShop(MWS) 2008. http://www.iwsec.org/mws/2008/en.html.
[2] Cuckoo sandbox. http://www.cuckoosandbox.org/.
[3] Cyber Clean Center. https://www.telecom-isac.jp/ccc/en_index.html.
[4] FFR yarai analyzer Professional. http://www.ffri.jp/products/yarai_analyzer_pro/.
[5] IIJ MITF. https://sect.iij.ad.jp/en/mitf.html.
[6] PRACTICE: Proactive Response Against Cyber-attacks Through International Collaborative Exchange. http://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/Releases/Telecommunications/130307_02.html.
[7] PREDICT, the Protected Repository for the Defense of Infrastructure Against Cyber Threats. https://www.predict.org/.
[8] WOMBAT project: Worldwide Observatory of Malicious Behaviors and Attack Threats. http://www.wombat-project.eu/.
[9] M. Akiyama, K. Aoki, Y. Kawakoya, M. Iwamura, and M. Itoh. Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks. *IEICE Trans. on Commun.*, E93-B(5):1131–1139, May 2010.
[10] K. Aoki, T. Yagi, M. Iwamura, and M. Itoh. Controlling malware HTTP communications in dynamic analysis system using search engine. In *Proceedings of Third International Workshop on Cyberspace Safety and Security*, pages 1–6, Sep. 2011.
[11] S. Miwa, T. Miyachi, M. Eto, M. Yoshizumi, and Y. Shinoda. Design and implementation of an isolated Sandbox with Mimetic Internet used to Analyze Malwares. In *Proceedings of DETER Community Workshop on Cyber Security Experimentation and Test*, pages 1–9, Aug. 2007.
[12] K. Nakao, D. Inoue, M. Eto, and K. Yoshioka. Practical Correlation Analysis between Scan and Malware Profiles against Zero-Day Attacks Based on Darknet Monitoring. *IEICE Trans. Inf. Syst.*, E92-D(5):787–798, May 2009.
[13] A. Nappa, M. Z. Rafique, and J. Caballero. Driving in the Cloud: An Analysis of Drive-by Download Operations and Abuse Reporting. In *Proceedings of the 10th Conference on DIMVA*, pages 1–20, Jul. 2013.
[14] Y. Zhou and X. Jiang. Dissecting Android Malware: Characterization and Evolution. In *Proceedings of 33rd IEEE Symposium on S&P*, pages 95–109, May 2012.