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

Mitsuhiro Hatada\textsuperscript{1}  
\textsuperscript{1}NTT Communications Corporation  
Gran Park Tower 16F, 3-4-1 Shibaura Minato  
Tokyo, Japan 108-8118  
m.hatada@ntt.com

Masato Terada\textsuperscript{1}  
\textsuperscript{1}Hitachi Incident Response Team  
1-1-2 Kashimada, Saiwai, Kawasaki, Kanagawa, Japan 212-8567  
masato.terada.rd@hitachi.com

Tatsuya Mori\textsuperscript{1}  
\textsuperscript{1}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.

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

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author(s). Copyright is held by the owner/author(s).

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 webclient, 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 yara 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)).

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-
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