Main Objective of MWS

Accelerate and expand the activities of anti-malware research by sharing attractive datasets with community.

Contributions

- To 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.
- To share the lessons learned from our experiences over the past seven years of sharing datasets with the research community – from the view point of Data and Lowering obstacles.

Data

- The packet traces have attracted the most newcomers for performing various analysis such as machine learning.
- The synchronization of the formats and collection periods of different datasets facilitates the identification of common and separate trends of attack.
- The types of datasets have been flexibly updated to remain abreast of threat transitions in the wild.

Lowering obstacles

- The technical obstacles of data collection such as developing and operating honeypots.
- The simple procedure for accessing these datasets as much as possible in order to to make the datasets available to any researcher wishing to conduct anti-malware research using datasets.
- The descriptions of the datasets in Japanese to avoid the neglect of students who are less capable with the English language.

Features of the datasets are applicable to several attack phases, assist researchers in performing the long-term analysis, and facilitate the correlation of various datasets collected by different research institutes and industries.

1) Probing: collected by darknet analysis
   - NICTER Darknet Dataset ('11~'14): packet traces collected from the darknet monitoring system, NICTER, which covers approximately 210 K unused IP addresses.

2) Infection: collected by server side and client side honeypot
   - CCC DATASET ('08~'13): list of hash digests for collected malware samples, packet traces, and the logs of malware collection collected from server-side, high-interaction distributed honeypots operated by the Cyber Clean Center.
   - IIJ MITF Dataset ('12): logs of malware collection from server-side, low-interaction distributed honeypots operated by MITF.
   - D3M ('10~'14): packet traces collected from web-client, high-interaction honeypot, Marionette and dynamic malware analysis system, Botnet Watcher.

3) Malware activities: collected by sandbox and forensic analysis
   - PRACTICE Dataset ('13): long-term packet traces collected from the dynamic malware analysis system operated by the PRACTICE project.
   - FFRI Dataset ('13~'14): logs collected from the dynamic malware analysis system Cuckoo sandbox and yarai analyzer Professional.
   - MARS for MWS ('08~'10): memory dump and forensic data collected from the dynamic malware analysis system using not-virtualized machine, MARS.

Seven Years of Experiences

Number of published papers related to malware in the largest domestic Computer Security Symposium in Japan.

MWS community growth (As of Jul. 12, 2014)

- The number of research groups tripled from '08 to '14.
- Roughly 30 groups constantly used of the datasets.
- New research groups have arisen every year.

- The new research groups: not worked in malware-related research in the past and their first paper on malware-related research was presented at MWS.

Number of published papers used MWS Datasets.

(As of Jul. 12, 2014)

- The total number of publications has reached 30 in the past five years.