WIDE Technical-Report in 2008

Meta-data format and associated tools for communicating PCAP analysis results

wide-tr-mawi-admd-00.pdf

WIDE Project : http://www.wide.ad.jp/

If you have any comments on this document, please contact to ad@wide.ad.jp

Title:	Meta-data format and associated tools for communicating PCAP analysis results
Author(s):	門林 雄基 (youki-k@is.aist-nara.ac.jp)
Date:	2008-12-24

Meta-data format and associated tools for communicating PCAP analysis results

Youki Kadobayashi
National Institute of Information and Communications Technology, Japan

1 Background

To date, many engineers and scientists have been working on PCAP files, yet we did not have any effective means to communicate what we have found. In other words, we are still in the dark ages of data analysis in this field, since the result of analysis cannot be communicated and compared with each other.

This is typically problematic in the cybersecurity context, since many scientists have been working on common datasets (e.g., the MAWI traffic archive) to locate anomalies, without being able to further validate their results with each other. Since real-world datasets do not have "correct class label" in most cases, relative comparison among multiple anomaly detection algorithms seems to be best alternative approach to improve their accuracy.

2 Common meta-data format for PCAP analysis

Here we consider adopting common meta-data format across different analysis techniques. If different analysis techniques can produce compatible mark-ups against the same dataset, we can compare their results without translating or converting the mark-ups.

There are lots of potential benefits that we can obtain from common meta-data format. More specifically, there are four kinds of direct beneficiaries, as described below.

Algorithm designers will benefit from the common meta-data format since their results will be made comparable among adopting parties. In addition, they will be freed from developing in-house data format to store the analysis result. Furthermore, they will benefit from additional tools built around the common meta-data format, e.g., tools for synthesizing datasets out of known anomalies and background traffic.

Cybersecurity researchers and practitioners will benefit from the meta-data format, because they will be able to benchmark multiple anomaly detection algorithms against the same dataset, without being involved in time-consuming data conversion process. In addition, they may choose to communicate their own analysis results in the same format, giving feedback to algorithm designers.

Tool implementers will benefit from existing common meta-data definition and associated class libraries. Also, they can test their newly developed tool against existing real data.

Dataset repository maintainers will benefit from common meta-data format, since it enriches the scientific value of shared dataset repository. The common meta-data format simplifies management of secondary data. It also helps analysts to document essential information for reproducible analysis; e.g., relationship of secondary data with original PCAP data, and parameters given to particular algorithm.

3 ADMD schema

As a starting point of meta-data format, XML Schema for annotating the result of analysis is made available[1], which we call ADMD (Anomaly Detection

[1] http://admd.sourceforge.net/

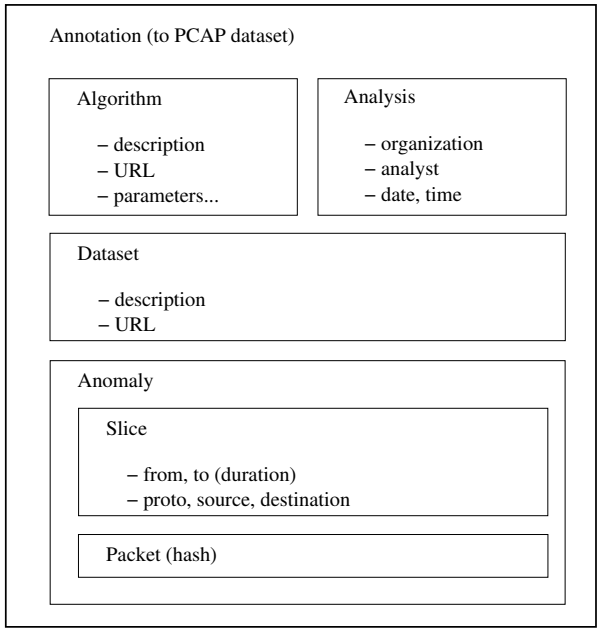

Figure 1: Outline of ADMD schema

Meta-Data), along with C and C++ API to annotate PCAP dataset according to the XML Schema. PCAP data analysis programs are supposed to use either C or C++ API to represent the result of analysis in the ADMD XML Schema. Data analysis programs written in other languages such as Java or Perl can also be supported through native-code wrappers.

The primary focus of this XML Schema is content (annotated results) and reproducibility (algorithm description and parameters). The envelope information of each PCAP dataset, e.g., date and observation point, should be better described by CAIDA's DatCat tools. This tool focuses more on individual record or flow in PCAP datasets.

The concise XML Schema currently consists of 8 data types, in 80 lines. The data types are organized in hierarchical manner, as depicted in Figure 1.

4 PCAP manipulation and validation tools

A set of toolchain is provided to 1) manipulate PCAP datasets according to mark-ups, and 2) compare anomaly detection results. They are described in the following.

`admd_slice` takes annotated result of analysis, represented in XML, and emits matching slice of the input PCAP file into the output PCAP file.

`admd_merge` takes annotated result of analysis, then injects matching slice of the second PCAP file into first PCAP file with the specified time offset, generating the output PCAP file.

`admd_validate` takes a pcap file and a set of annotated analysis results in XML. It is intended to compare the performance of variety of algorithms.

5 Next steps

We have been working with algorithm designers to improve the proposed ADMD schema and toolchain. We are looking forward to see more scientists, who will benefit from public PCAP dataset and existing secondary datasets that are created through ADMD.

We are also looking into collaboration with cybersecurity researchers and practitioners by developing more operator-friendly interfaces. We already have minimal, Eclipse-based environment for editing ADMD-compliant annotations.

In near future, we will have to work with Dataset repository maintainers for general issues pertaining to archival of secondary data, e.g., naming conventions.