ADFA-IDS DATASETS COMPOSITION AND UTILIZATION

This document describes the general description, composition and how to use the following ADFA-IDS datasets:

(1) **NGIDS-DS dataset:-**

This dataset is generated at the next generation cyber range infrastructure of the Australian Centre OF Cyber Security (ACCS) in the University of New South Wale (UNSW)@ Australian Defence Force Academy(ADFA), Canberra. It is the part of the ongoing projects in the ADFA related to the cyber security.

It is the collection of normal and abnormal host (LINUX) and network activities which are performed during the emulation. It contains four main type of files for the evaluation of future IDS design evaluation. The first type of files are considered as host log files, which are provided in the folder host logs. In host logs folder, there are 99 csv files. The second type of file is considered as network log file and its name is NGIDS.pcap. The third file contains the ground truth information and its name is ground_truth.csv. The fourth file name is feature_descr.csv, it holds the attributes information for the host logs and ground truth information. More details description of NGIDS-DS is available in the article with title *"Generating Realistic Intrusion Detection System Dataset based on Fuzzy Qualitative Modeling"*.

In the case of Host based Intrusion/Anomaly Detection System (HIDS/HADS) design evaluation, the host logs of NGIDS-DS can be used. In order to evaluate Network Intrusion/Anomaly Detection System (NIDS/NADS) design NGIDS-DS pcaps can be used. Further, to evaluate the design of combined IDS, both host logs and pcaps can be used. Moreover, the general details, regarding how to compose training and test data for signature or anomaly IDS cases and how to measure accuracy and error, are provided under the datasets 2 and 3.

(2, 3) ADFA-LD and ADFA-WD datasets

ADFA-LD, ADFA-WD and ADFA-WDSAA are labelled data that contains following three different
folders:

(i) Training data (contains only normal traces)
(ii) Validation data (contains only normal traces)
(iii) Attack data (contains only attack traces)

In the case of anomaly intrusion detection system (IDS) design you can use just training normal data in training phase with normal label if you want to do supervised ML and at testing phase you can use all attack data and validation data to measure Detection rate (DR), False positive rate (FPR), False negative rate (FNR) and False alarm rate (FAR). In the case of signature IDS design you can use all normal training data with label normal and some attacked data (from test attack data) with label attack while training phase and using supervised machine learning (ML). During testing you can use rest of the attack data that is not use in training and all validation normal data, to measure DR, FPR, FNR and FAR.In the case of unsupervised ML specific to anomaly IDS design, the training phase would require normal training data without labels and for testing phase, all the attack data and validation data with labels can be used to measure DR, FPR, FNR and FAR. Further the audience can use the following research articles for study and reference as they utilized the abovementioned datasets.

(a). Haider, Waqas, Jiankun Hu, and Miao Xie. "Towards reliable data feature retrieval and decision engine in host-based anomaly detection systems." IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), 2015.

(b). Haider, Waqas, et al. "Integer Data Zero-Watermark Assisted System Calls Abstraction and Normalization for Host Based Anomaly Detection Systems." IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), 2015.

(c). Haider, Waqas, et al. "Windows Based Data Sets for Evaluation of Robustness of Host Based Intrusion Detection Systems (IDS) to Zero-Day and Stealth Attacks." Future Internet 8.3 (2016): 29.

(4) netflow_ids_label dataset

This dataset is specifically generated for NIDS/NADS evaluation. It contains the labbled network flow information. More specific details are available in the article with the title "*A real-time NetFlow-based intrusion detection system with improved BBNN and high-frequency field programmable gate arrays".*

Note: NGIDS-DS host logs and netflow IDS dataset, are labelled via column with entries 0(normal) and 1(abnormal), whereas ADFA-LD and ADFA-WD are labelled via folders with names (e.g. Normal data, Attack data, and

Validation data). Therefore, the user can make the particular training and testing composition while considering the label information of the data files.