Network Security and Measurement Assignment 03

HAW Hamburg WS 2021

Prof. Dr. Thomas Schmidt, Raphael Hiesgen, M.Sc. Deadline: November 17, 2021

When approaching large datasets it is helpful to preprocess data to simplify analysis later on. Read through all three exercises before starting. Even though we will only process data for one day (due to time constrains) consider that for a representative analysis it might be necessary to analyze one month or even one year of data.

1. **WIDE Traffic Repository**

   WIDE is a project that runs major parts of the Internet in Japan. It provides several traffic traces (`https://mawi.wide.ad.jp/mawi/`), which we want to study. For privacy reasons, the dumps do not include the application header (*i.e.,* payload of the transport header). Furthermore, all IP addresses are anonymized, but the anonymization preserves the original IP prefix.

   Data: MAWI data is located in `shared-data/mawi`. Use the first 15-minute file for November 1st. It is cut into minute-long slices in `shared-data/haw/mawi/2021/in-minutes`.

   Tools: tshark, dpkt, scapy, pandas, ...

   (a) Calculate and visualize the number of bytes and packets per minute.

   (b) Calculate and visualize the distribution of packets and bytes per flow.

2. **Port-based Classification**

   In most cases ports tell a lot about the mix of services and applications in use. What port do you expect to see most?

   (a) Implement a script to classify packets based on their ports.

   (b) Give an overview of the services in use for each layer (network layer, transport layer, and potential application layer). Rank them by popularity in terms of bytes and packets.

   (c) Compare your results to the findings in *Distilling the Internets Application Mix from Packet-Sampled Traffic* (Richter et al., IMC'13).

3. **Client vs. Server**

   Throughout the Internet most communication happens between clients and servers. Distinguishing between them gives a clearer view of the traffic and helps understand specific phenomenons that might only affect one them (*e.g.,* reachability and traffic mix).

   (a) Design an approach to group the data by server and clients.

   (b) Implement your approach, show results, and discuss limitations.