LaSIC - Labeled Security Information Capture

Developing robust security for a modern network requires high-quality, trustworthy datasets to tune, test and validate security models and mechanisms. Often this requires labeled datasets, which are extremely hard to obtain by the security community. Labeled datasets are essential to the machine learning (ML) community, but also important to all security research to help establish ground truth.

LaSIC addresses this problem by capturing and labeling real traffic from a production network that serves several science communities. By providing labeled datasets to security researchers, LaSIC increases the understanding of operational and/or realistic scientific cyberinfrastructures, helps in developing new and novel cybersecurity technologies, and provides real data for rigorous and realistic testing, evaluation, and validation of cybersecurity research. AmLight is unique in terms of the diversity of the scientific community it serves (Climate, High-Energy Physics, Astronomy, Genomics), the geography of the network and its customers (North and South America, Africa and Europe), the unique hardware (100G fiber terrestrial and underwater links, Internet Telemetry-capable switches, high-end routers, custom unsampled packet header and flow capture tools), and allowing multimode data capture (packet headers, NetFlow, In-band Network Telemetry, intrusion detection systems, SNMP) from the same links. Each of these modes has its own advantages: unsampled packet headers and flows are needed to discover needle-in-a-haystack phenomena; NetFlow and SNMP summarize traffic making datasets smaller and easier to handle, but still capture large phenomena such as DDoS attacks; INT data reveals micro-phenomena such as microbursts; and IDS alerts provide external DDoS labeling, which can be hard and/or tedious to do.

LaSIC Datasets

Dataset Organization Labels. These are typically chosen to provide basic information about the file contents, are relatively easy for humans or simple scripts to interpret, and facilitate searching. Examples of informative filename components include a date and timestamp, a string representing the location of capture (e.g., routerA or subnetB), a sequence number for multifile datasets, and an extension representing the format of the data (e.g., .csv, .pcap, .nf, etc).

Security Event Labels. We define a security event label as metadata used to describe a security event captured in the raw data. Examples include microbursts (short periods of very high network activity), DDoS attacks as determined by an IDS configured by network operators, a high rate of authentication failures as determined by numerous failed login attempts, and so forth.

Network Structure Labels. These labels include details about the structure of the network. Such labels include (a) network layer protocols (IPv4 vs. IPv6, TCP vs. UDP, etc.); (b) Internet services (DNS, Web, NTP, SSH, RPC, etc.); (c) IP addresses for servers providing standard Internet services; and (d) custom services used by scientific research groups, such as repositories, FTP servers etc., that we will discover through scanning or consulting with local network operators.

We will present LaSIC and the datasets it currently captures, labels and distributes to the community. LaSIC partners with Classnet to advertise the datasets. Classnet provides a searchable catalog of LaSIC datasets, as well as the handling and management of dataset requests.