Labeling Network Telescope Data: Challenges and New Directions (Abstract)

Michalis Kallitsis
mgkallit@merit.edu

Network telescopes consist of networking infrastructure that record and receive unsolicited Internet-wide activities. A network telescope or “darknet” is configured to listen to traffic destined to an unused but routed IP space. Since this “dark IP space” serves no legitimate network services such as Web servers, DNS resolvers, etc., any traffic arriving to the darknet is inherently suspicious/malicious. Thus, network and security researchers have been utilizing network telescopes to study a plethora of macroscopic Internet activities such as shedding light into botnets [1], [2], obtaining insights about network outages [3], [4], understanding certain types of denial of service attacks [5], [6], [7], examining the behavior of IoT devices [8], and detecting Internet misconfigurations [9], [10], etc.

Despite the indisputable utility of network telescopes, annotations and labeling of darknet activities usually happen on an opportunistic basis (e.g., to study, in a post mortem manner, a known event, such as the onset of the Mirai botnet [1]). The problem of automated and ongoing labeling has been a challenging one, exacerbated by the sheer volume of darknet data that the analyst needs to consider. In this presentation, we would like to describe recent efforts on renewing Merit Network’s network telescope and adding meaningful annotations on the collected data. Specifically, we would like to present the ORION (Observatory for cyber-Risk Insights and Outages of Networks) infrastructure [11] and its data pipeline that extracts Darknet events of interest (such as scanning activities and spoofing-based denial of service attacks) while also uploading—in near-real-time—such annotated data into Google’s BigQuery for ease of processing and analysis. The ORION data pipeline enriches the identified Darknet events with several other useful meta-data (such as routing, DNS and geolocation information) along with useful fingerprints/labels that can be extracted from packet headers (i.e., the Mirai, Masscan and ZMap fingerprints).

The ORION network telescope has also enabled new opportunities for data labeling. We would like to discuss our work on clustering darknet data using AI/ML techniques such as unsupervised clustering [12]. In this work, our focus is on assigning network-based features to scanning IPs appearing in the ORION network telescope, and using unsupervised learning techniques, such as K-means, to assign the darknet scanners into groups of similar activities. To deal with the high-dimensionality of the input feature space, we employ “deep representation learning” techniques to “compress” the input data into a lower-dimensional space of embeddings that we would then use for clustering. We will discuss how the clustering output can be leveraged for the detection of temporal changes in the structure of the darknet (i.e., finding when “new” activities may be appearing in the darknet), thus enabling novel ways to offer meaningful threat intelligence to the cybersecurity community.
References


