Community Labeling and Sharing of Security and Networking Test datasets

Jelena Mirkovic[1], John Heidemann[1], Wesley Hardaker[1], and Robert Stovall[2]

[1]USC Information Sciences Institute
[2]Merit Networks, Inc.

1 Abstract

Community Labeling and Sharing of Security and Networking Test datasets (CLASSNET) project aims to support network and security research with new, labeled, rich and diverse datasets to the research community. The project is developing a framework for collaborative, community-driven enrichment and labeling of data, enabling use of our datasets for machine learning in networking and security. Second, the CLASSNET project makes data available to researchers through multiple methods, ensuring privacy of data while enabling flexible data computation. Finally, the project also generates diverse continuous (constantly, automatically updated) and curated (selected by human) datasets for research use.

In addition to providing data to the research community, CLASSNET project aims to innovate in dimensions of data labeling, data distribution and data sources. For data labeling CLASSNET provides a collaborative framework via for low-friction sharing of annotations among researchers. The framework supports bulk, automatic, algorithmic labeling. For data distribution, CLASSNET supports multiple ways of data access, ranging from downloading anonymized data to processing data in cloud, on provider machines or via the code-to-data approach. Finally, CLASSNET data sources provide new, diverse, continuous, and curated datasets that are useful for network and security research, including traffic packets and flows, network telescope data, DNS data and Internet topology data.

CLASSNET data is currently available to researchers via our COMUNDA portal at https://comunda.isi.edu. Researchers can log in using their Google, Github or institutional credentials. They can browse and search the datasets and submit online requests for any dataset of interest, including signing of provider-specific data use agreements. Requests are manually processed by data providers. Once approved, the researcher receives an email with details of how to access the datasets. Researchers can also comment on and rate datasets they have accessed. Each dataset can be labeled by researchers, by following a process outlined in the documentation of the COMUNDA portal.

CLASSNET project currently supports three data providers: USC/ISI, Merit Networks, and University of Memphis. We are currently building mechanisms to support a larger data provider community via our COMUNDA portal. Providers can contribute flexibly to the COMUNDA portal. At the minimum a provider contributes metadata of their datasets, and keeps ownership of the data and the approval process. As a researcher requests a dataset, our systems generate an automated email to the provider with the data use agreement, which the researcher has filled out online and signed electronically. The provider can approve or deny access by accessing our ticketing system. CLASSNET project will further support providers that want to offload data hosting, access approval or both to the CLASSNET staff.

With the Internet's importance for tele-work, tele-medicine, remote learning, e-commerce and e-government, robust network and security are of paramount importance. We hope that the broader impact of this project will be to foster wider sharing and labeling of networking and security datasets, and aid research and education in these scientific fields. CLASSNET project's innovations in multiple pathways to data access, combined with the automated and incentivized enrichment framework, will improve the state-of-the-art for responsible data sharing in related disciplines of information technology.