An Empirical Analysis of the Impact of Data Sharing on Anomaly Detection for Network Security

Authors

Abstract

Anomaly detection is used to discover and understand network security issues, especially pertaining to distinguishing between normal and attack traffic. Since a single entity has a limited vantage point from which to observe network traffic, the sharing of network traffic data has been proposed as a method to improve the performance of network anomaly detection. Existing work has largely focused on sharing characteristics of detected anomalies (Arjun:Check this). In this paper, we take a broader view of network data sharing and aim to characterize when it can improve anomaly detection performance. In particular, we focus on the sharing of network traffic data and the models trained on them in the supervised, unsupervised and semi-supervised learning paradigms. For each mode of learning, we consider the impact of the volume of data, its heterogenity and labeling errors on the performance of anomaly detectors. Our experiments on 2 publicly available and 1 curated dataset, using 5 different types of models indicate that: TODO:add experimental conclusions here

1 Introduction

- Key questions about sharing of benign data alone: - If homogeneous, more data will not hinder performance?
- If heterogeneous, expressive enough (deep) models should be able to learn, what happens with shallow models?
- Questions about malicious data sharing: - Not shared for unsupervised - Shared along with labels for S and SS

Contributions: 1. Clearly map out a taxonomy of data sharing for network security 2. Establish baselines when data sharing is helpful (with regards to data, model and sharing) 3. Practical method to model and understand data heterogenity

- What's missing? Different modes of data sharing, different learning paradigms, unclear what the performance gains from sharing are
- This paper characterizes in what scenarios data sharing can help: - Data heterogenity when building a model of 'normal behavior' (affects FP rates, will not affect TP)
- Modeling data het. is an open problem, especially for different classes of benign data - Different learning paradigms: how is unsupervised learning aided by more data?
- Robustness to errors in labeling (ending up with malicious data in benign split)

- Experiment structure: - Datasets: IDS, IoT, BAS - Models: RF, IF - Direct data sharing: S, US, SS; tune data het. for each - Model sharing: S, US, SS - Interpretation of results for different attack types

2 Related Work

Foo[**reich2013modular**].

2.1 Learning paradigms for network data

2.2 Data sharing for cybersecurity

2.3 Collaborative learning

3 Conclusion