Detecting Temporal and Spatial Anomalies in Users' Activities for Security Provisioning in Computer Networks

Aleks Huč
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
aleks.huc@fri.uni-lj.si

Thesis Summary

Keywords: anomaly detection, incremental learning, unsupervised learning, clustering, adaptive windowing, profiling, network security, network flows

Received: June 1, 2023

The paper summarizes a Doctoral Thesis that focuses on two new approaches for detecting anomalies in computer networks based on network flows. The approaches use incremental hierarchical clustering algorithms and monitor changes in the data structures to detect anomalies. Both approaches achieved prediction performance comparable to the state-of-the-art supervised approaches ($F1$ score over 0.90), even when taking into account that our approaches see every data point only once and then discard it and they operate without the prerequisite learning phase with labeled data.

Povzetek: Članek povzema vsebino doktorske disertacije, v kateri se osredotočimo na dva nova pristopa za detekcijo anomalij v računalniških omrežjih. Pristopa temeljita na omrežnih tokovih, inkrementalnem hierarhičnem gručenju in spremljanju sprememb v podatkovnih strukturah z namenom detekcije anomalij. Oba pristopa doseženta primerljivo stopnjo detekcije (mera $F1$ preko 0.90) v primerjavi z najnovejšimi nadzorovanimi metodami, tudi ko upoštevamo, da naša pristopa vidita vsak podatek le enkrat in ga nato pozabita ter delujeta brez predhodne faze učenja z označenimi podatki.

1 Introduction

The goal of computer network security is to provide a secure environment for a computer network, its resources, data in storage and transit and all its users [1]. Network security starts with intrusion detection, which is defined as a deliberate unauthorized attempt (successful or not) by an intruder to gain access to, manipulate or misuse a computer system or network [1]. Examples include Trojans, viruses, malware and denial of service, brute force and probe attacks.

Over the years of active development, two main categories of intrusion detection approaches have emerged: signature-based and anomaly-based [2]. Signature-based approaches detect intrusions on the basis of signature databases while anomaly-based approaches detect intrusions on the basis of deviations from normal activity models. Various anomaly detection approaches have already been proposed but have problems with today's dynamic computer networks with large volume and high velocity, variety and variability due to their use of supervised and batch learning. Newer methods have switched to unsupervised, incremental and adaptable methods to improve upon and augment traditional approaches and provide overall better anomaly detection.

This paper summarizes a Doctoral Thesis [3] that provides two new approaches for improving the current state-of-the-art anomaly detection using unsupervised, incremental, adaptable and hierarchical clustering.

2 PHICAD

PHICAD (Profile- and Hierarchical Incremental Clustering-based Anomaly Detection) is a single-layer, unsupervised and incremental algorithm that detects network activity anomalies in real-time. The input is a stream of chronologically ordered flows. The algorithm receives a new flow and sends it to the appropriate two profiles based on source and destination IP addresses. A profile models the incoming and outgoing activity of an individual network entity. The algorithm then extracts, transforms, and normalizes the features from the flow into a real-valued vector. The vector is then clustered inside the appropriate profile hierarchical clustering tree structure.

The anomalies are determined in the leaf nodes where, if the new vector is merged with the existing leaf, we track the distance between the new vector and the leaf, the leaf centroid changes and the leaf size; or if the new vector becomes a new leaf, we track the distance between the new leaf and the centroid of neighboring leaves. The predictions from all detection mechanisms are put into a short-term model that discards mechanisms that trigger too often and reports final predictions.

3 PHI2CAD

PHI2CAD (Profile- and Hierarchical Incremental Two-layer Clustering-based Anomaly Detection) builds upon our single-layer PHICAD with an additional second layer unsupervised and incremental clustering algorithm which detects anomalies in profiles and groups of profiles.

The input into our approach is again a stream of chronologically ordered flows. First, the flow is sent to the first layer where the PHICAD algorithm creates and updates profiles of network entities and detects network anomalies in each individual profile separately. For each flow, the PHICAD produces two updated profiles with predictions for possible anomalies, one for the source and one for the destination IP address, which are then sent to the PHI2CAD algorithm on the second layer.

PHI2CAD first checks for each updated profile if it has already been clustered into its tree data structure and if it has been, it checks if the updated profile is still inside the leaf or not. If it is still inside, we check for possible anomalies caused by the updated profile and produce possible anomaly predictions, by tracking the distance between the updated profile and the leaf, the leaf centroid changes, and the leaf size.

Otherwise, if an updated profile has not been clustered yet or it falls outside the leaf it has previously been clustered to, we cluster the updated profile into PHI2CAD tree data structure, while its previous version, if it exists, is removed from the tree. Finally, possible anomaly predictions are determined in the leaf to which the updated profile has been clustered. If the updated profile is merged with the existing leaf, we track the distance between the updated profile and the leaf, the leaf centroid changes and the leaf size; or if the updated profile becomes a new leaf, we track the distance between the new leaf and the centroid of the neighboring leaves. The predictions from all detection mechanisms are input into a short-term model that discards mechanisms which trigger too often and reports the final predictions.

4 Conclusion

The goal of this dissertation was to research if we can devise an anomaly detection approach with the following operational constraints: incremental execution, unsupervised learning, real-time response, ability to analyze large data sets, lightweight design and ability to adapt to changes over time, while still providing comparable performance to classic approaches and/or providing us with additional new insights.

We have evaluated our two approaches using a state-of-the-art data set CICIDS2017 [4] that comprises the most common network anomalies. To measure the predictive performance we used standard machine learning metrics such as precision, recall and F1 score and also the execution time against the supervised approaches. To further explain the achieved prediction performance we analyzed the influence of individual features on the predictions and performed sensitivity analysis of the main parameters.

Our approaches can successfully detect Denial of Service, Distributed Denial of Service, Port Scan and Web attacks when analyzing each anomaly separately and are also able to detect anomalies even when they analyze entire data sets with multiple types of anomalies. Performance is good where anomalous patterns clearly differ from the normal activity (F1 score over 0.90), however, they have problems detecting attacks that are presented with flows that are similar to normal flows or that are executed on higher layers of the network stack or are a part of packet payloads. But we have to be mindful of the diminishing importance of packet-payload analysis, due to the increasing use of packet-payload encryption. The results were also published in a peer-reviewed journal paper [5].

References

[1] Kizza, J. M. (2020) *Guide to computer network security*, Springer.

[2] Thakkar, A. and Lohiya, R. (2021) A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions, *Artificial Intelligence Review*, Springer, pp. 1–111.

[3] Huč, A. (2022) *Detecting temporal and spatial anomalies in users' activities for security provisioning in computer networks*, doktorska disertacija, Ljubljana, https://repozitorij.uni-lj.si/IzpisGradiva.php?id=137562.

[4] Sharafaldin, I. and Lashkari, A. H. and Ghorbani, A. A. (2018) Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterizationy, *4th International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108–116.

[5] Huč, A. and Trček, D. (2021) Anomaly detection in IoT networks: From architectures to machine learning transparency, *IEEE Access*, IEEE, pp. 60607–60616.