DetGen: Deterministic Ground Truth Traffic Generation using Docker for Machine Learning

Anonymous Author(s)

ABSTRACT

ACM Reference Format:
Anonymous Author(s). 2018. DetGen: Deterministic Ground Truth Traffic Generation using Docker for Machine Learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The exponential growth of data availibity enabled the machine learning revolution of this decade and transformed many areas of our lives. Ironically, researchers struggle to gather qualitative network traffic data to ... Well-designed datasets are such a rarity that researchers often evaluate intrusion detection systems on datasets that are well over a decade old [**?** **?**], calling into question their effectiveness on modern traffic and attacks. The lack of quantity, variability, meaningful labels, and ground truth has so far prohibited ML-based methods from having a bigger impact in network security.

Privacy and security concerns discourage network administrators to release rich and realistic datasets for the public. Network traffic produced by individuals contains a host of sensitive, personal information, such as passwords, email addresses, or usage habits, requiring researchers to expend time anonymising the dataset [**?**]. In order to examine malicious behaviour, researchers are often forced to build artificial datasets using isolated machines in a laboratory setting to avoid damaging operational devices. Background traffic is generated either from ...

The datasets currently available are meant to be all-purpose and are *static* in their design, unable to be modified or expanded. This proves to be a serious defect as the ecosystem of intrusions is continually evolving. Furthermore, it prohibits a more detailed analysis of specific areas of network traffic. To prevent this, new datasets must be periodically built from scratch.

Additionally, ground truth

Developing a framework that allows researchers to create datasets that circumvent these issues would be extremely beneficial. We propose that this can be done using Docker [**?**]. Docker is a service for developing and monitoring containers, also known as OS-level virtual machines. Each Docker container is highly specialised in its purpose, generating traffic related to only a single application process. Therefore, by scripting a variety of Docker-based *scenarios* that simulate benign or malicious behaviours and collecting the resultant traffic, we can build a dataset with perfect ground truth. Furthermore, these scenarios could be continually enhanced and expanded, allowing for the easy creation of datasets containing modern, up-to-date traffic and attacks.

This is the primary goal of this work. Furthermore, we demonstrate the utility of this framework by performing a series of experiments: one that measures the realism of the network traffic produced by our Docker scenarios and two that would be difficult to perform using a conventional dataset.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 ACKNOWLEDGMENTS

We are grateful for our ongoing collaboration with our industry partners (blinded) on this topic area, who provided both ongoing support and guidance to this work. Discussions with them have helped reinforce the need for a better evaluation and understanding of the possibilities that new intelligent tools can provide.

Full funding sources after currently blinded.

A RESULT TABLES