

Thesis  
M.Sc.

IDP

## Creation of Data Sets in Automotive Environments for Network Anomaly Detection

### Motivation

The most important enabler for machine learning is the underlying (training) data. However, this data has to reflect the processes of the modeled system in a realistic way. Otherwise, the model trained with this data would be incorrect.

A domain which relies more and more on models learned with machine learning is the automotive industry. Here, machine learning is used for different use cases, e.g. autonomous driving, predictive maintenance, monitoring or optimization. In this thesis, we want to create a training data set suitable for machine learning for network anomaly detection. This is an important goal, as most current publications either rely on very old data [1], or publicly inaccessible or synthetic data that is created for this sole purpose. As there is no common ground for evaluating the performance of the implemented methods, it is hard to compare them. Therefore it is desirable to have new, up to date data sets that can be used as reference and evaluation data sets in future research.

### Topic

The goal of this thesis is the creation of representative data sets in the automotive environment that can be used for future research. Therefore it is necessary to find out how a representative network capture in such an environment looks like. It is important to consider how and where such data can be collected. Finally it needs to be assessed in which regards this data set represents a valid collection that can be used in future research. An according pipeline needs to be designed and evaluated such that new data-sets can be created easily in the future.

### Your Task

- Assessment of available data sets
- Assessment of possibilities to collect data
- Identification of important data features and properties e.g., Connected entities, Communication relationships, etc
- Design and implementation of a pipeline for data set creation, resulting in a tool chain for quick data creation

### Requirements

- Knowledge in Python
- Basic network knowledge
- Ability to write maintainable code

### Sources

- [1] 1999 DARPA Intrusion Detection Evaluation Dataset, <https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset>

### Contact

Lars Wüstrich      wuestrich@net.in.tum.de  
Christian Lübben      luebben@net.in.tum.de  
Holger Kinkelin      kinkelin@net.in.tum.de

<http://go.tum.de/080755>

