Anomaly Detection in ISP Networks

Josef Koumar[1,2]

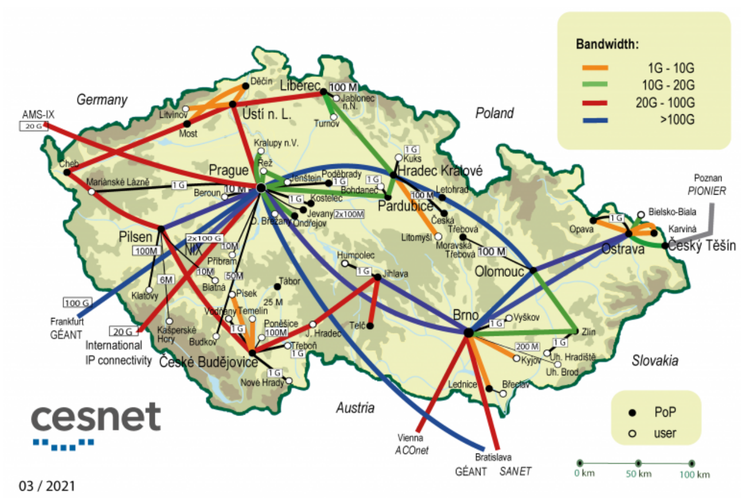
[1]Czech Technical University in Prague, Czech Republic; [2] CESNET, a.l.e.

Motivation

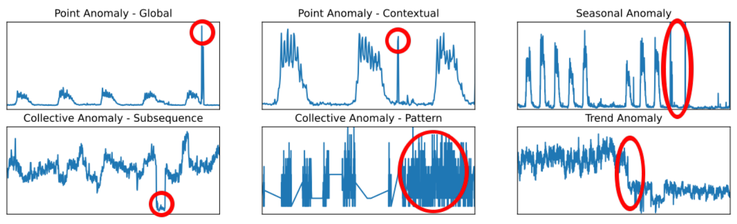
In recent surveys, the lack of a reference dataset for network traffic forecasting and anomaly detection is described as the crucial obstacle related to performance evaluation. Additionally, real-world datasets used in the evaluation are not publicly available due to privacy concerns.

Dataset creation

We decided to create a dataset called **CESNET-TimeSeries24** that was collected by long-term monitoring of selected statistical metrics for **40 weeks** for each IP address on the ISP network CESNET3 (Czech Education and Science Network).



The dataset encompasses network traffic from more than 275,000 active IP addresses, assigned to a wide variety of devices, including office computers, NATs, servers, WiFi routers, honeypots, and video-game consoles found in dormitories. Moreover, the dataset is also rich in network anomaly types since it contains all types of anomalies, ensuring a comprehensive evaluation of anomaly detection methods.



Last but not least, the CESNET-TimeSeries24 dataset provides traffic time series on institutional and IP subnet levels to cover all possible anomaly detection or forecasting scopes. Overall, the time series dataset was created from the **66 billion IP flows** that contain **4 trillion packets** that carry approximately **3.7 petabytes of data**. The CESNET-TimeSeries24 dataset is a complex real-world dataset that will finally bring insights into the evaluation of forecasting models in real-world environments. Download the dataset now:

This research was funded by the Ministry of Interior of the Czech Republic, grant No. VJ02010024: Flow-Based Encrypted Traffic Analysis and also by the Grant Agency of the CTU in Prague, grant No. SGS23/207/OHK3/3T/18 funded by the MEYS of the Czech Republic.

Dataset analysis

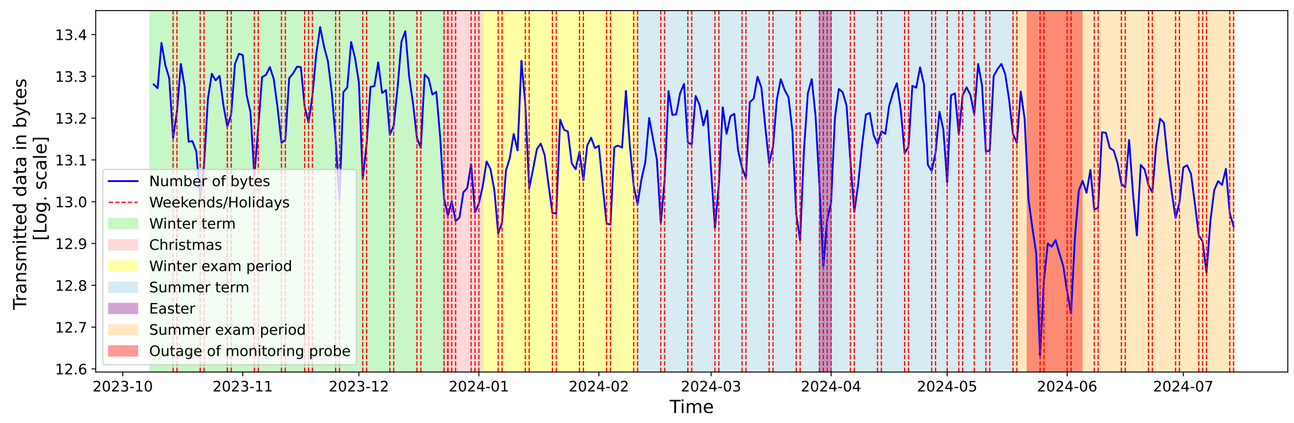
The dataset contains time series for several identifiers:

- IP addresses (original)
- Institutions
- Institution subnets

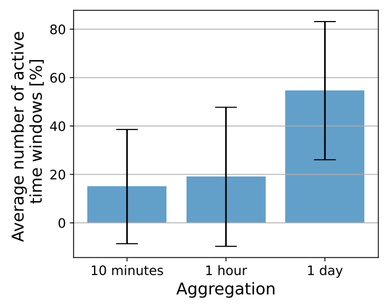
The dataset contains time series aggregated per several aggregation intervals:

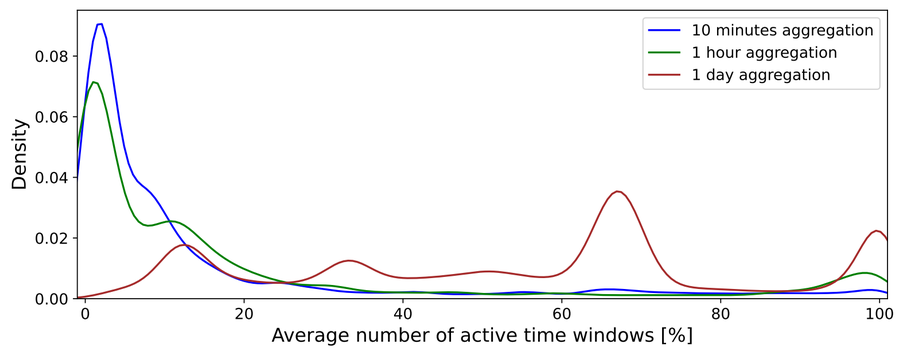
- 10 minutes (original)
- One hour
- One day

Overall data that were transmitted on the CESNET3 network and are captured in the dataset:



Analysis of gaps in dataset:

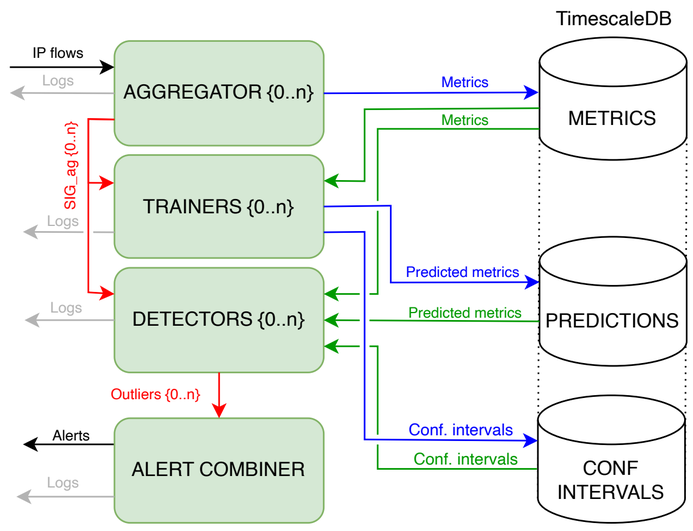




Anomaly Detection System

We propose a novel modular Network Outlier Detection System (NODS), which is built from open-source software listed in the table below. The NODS enables the deployment of outlier detection methods based on the forecasting of network traffic in real-world scenarios. We **successfully deployed** the system on the **real-world ISP network CESNET3**, where the system was in the testing phase for several months to ensure stability and reliability of the system.

Name	GitHub
ipfixprobe	CESNET/ipfixprobe
IPFIXcol2	CESNET/ipfixcol2
NEMEA Framework	CESNET/Nemea-Framework
NEMEA modules	CESNET/Nemea-Modules
NEMEA Supervisor	CESNET/Nemea-Supervisor
TimeScaleDB-14	timescale/timescaledb



Open Challenges

1. Many forecasting models can be computationally intensive, making it difficult to scale them for large datasets or in real-time applications. Optimizing models for efficiency while maintaining accuracy is crucial.
2. Many advanced forecasting models, especially deep learning-based ones, can be seen as black boxes. Providing explanations for why an anomaly was detected is critical for gaining trust from stakeholders and enabling informed decision-making.
3. Alerts should not only indicate that an anomaly has occurred but also provide context. Understanding the potential causes and implications of an anomaly is essential for effective response and mitigation.
4. In real-world systems, multiple anomalies can occur simultaneously. Developing methods to assess the correlation between different alerts and understand their combined impact is crucial for prioritizing responses.

Contact: koumajos@fit.cvut.cz