Poster: Learning distributions to detect anomalies using all the network traffic

Poster: Learning distributions to detect anomalies using *all* the network traffic

Alexander Dietmüller
ETH Zürich
adietmue@ethz.ch

Georgia Fragkouli
ETH Zürich
gfragkouli@ethz.ch

Laurent Vanbever
ETH Zürich
lvanbever@ethz.ch

ABSTRACT

Anomaly detection is an essential building block of many applications, including DDoS detection, root cause analysis, traffic estimation, and change detection. A vital part of detecting anomalies is establishing a sense of normality, e.g., by learning distributions for various features from benign traffic. Learning these distributions in the control plane requires coping with the limited visibility of sampling; learning distributions in the data plane requires relying on simplistic techniques because of hardware constraints.

We propose a novel data- and control-plane co-design for learning distributions: in the control plane, we search for candidate distributions with Bayesian optimization; in the data plane, we evaluate how well each distribution matches *all observed* traffic, without missing rare events. The aggregated evaluation results are fed back to the control plane to guide the optimization and learn accurate distributions. Our key insight is that while learning and optimization are infeasible in the data plane, evaluating distributions is feasible and leverages data plane strengths. We confirm the feasibility of our approach with a preliminary evaluation.

CCS CONCEPTS

• **Networks** → **Programmable networks**; **Network monitoring**;
• **Computing methodologies** → *Anomaly detection.*

KEYWORDS

Anomaly Detection, Traffic Distribution Learning, Bayesian Optimization, Programmable Networks, Network Monitoring.

ACM Reference Format:
Alexander Dietmüller, Georgia Fragkouli, and Laurent Vanbever. 2023. Poster: Learning distributions to detect anomalies using *all* the network traffic . In *ACM SIGCOMM 2023 Conference (ACM SIGCOMM '23), September 10, 2023, New York, NY, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3603269.3610837

1 INTRODUCTION

While essential to DDoS detection, root cause analysis, and many more applications, detecting anomalies in network traffic is notoriously hard to get right. Anomaly detection requires establishing a sense of normality [2]. This includes *rare* events (e.g., tails of traffic distributions) to avoid errors on uncommon yet benign traffic.

ACM SIGCOMM '23, September 10, 2023, New York, NY, USA

ACM ISBN 979-8-4007-0236-5/23/09…$15.00
https://doi.org/10.1145/3603269.3610837

Current approaches struggle to capture these events. The control plane can run arbitrarily complex algorithms but has limited visibility into the traffic (even sampling rates as high as 30% are not enough [4]). In contrast, a programmable data plane can observe all traffic but only run simple algorithms. For example, ACC-Turbo [1], a state-of-the-art detection system in the data plane, uses an online-learning approach that adaptively clusters traffic. However, it can only track the minimum and maximum values of traffic features,[1] and cannot track the distribution within a cluster. It identifies anomalies as high-volume clusters with high similarity, i.e., similar min and max values. As such, it is vulnerable to rare outliers that stretch the min/max value of a cluster, making an anomalous cluster seem dissimilar and harmless, even if the distribution remains narrow.

We argue that anomaly detection can be improved by learning feature *(i) distributions* based on the *(ii) entire* traffic. Is that feasible given that neither the control- nor the data plane achieves both?

We show that this is indeed possible through a data- and control-plane co-design that combines the strengths of both: Our key insight is that learning distributions is only feasible in the control plane, but once we have a distribution, it is possible to verify it in the data plane. This verification can leverage the data-plane visibility of all traffic without being limited by sampling.

In particular, we combine Bayesian optimization (BO) in the control- with scoring in the data plane: Searching optimal distribution parameters via BO is complex and must run in the control plane. Yet, for each proposed distribution, BO requires only an 'objective' value to advance the search. We use the logarithmic score as objective, and compute it from all traffic in the data plane. This forms a feedback loop of optimization in the control- and scoring in the data plane, iteratively learning distributions. Furthermore, we show that we can use scoring to detect anomalies, as such traffic scores significantly differently from expected traffic.

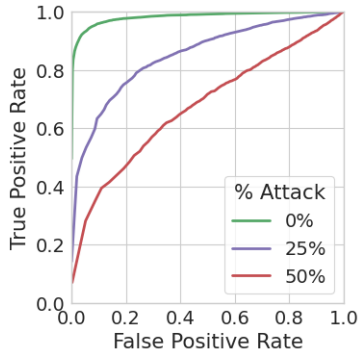
2 BAYESIAN SCORE OPTIMIZATION

Traffic features follow an arbitrary, unknown distribution P. We use Bayesian optimization (BO) to find a parametrized distribution Q to approximate P. We cannot directly measure how well Q matches P, but we can score Q based on observed packets. BO allows for maximizing a black-box function, and by maximizing a *proper scoring rule*, we guarantee that the optimal Q minimizes the distance to P.

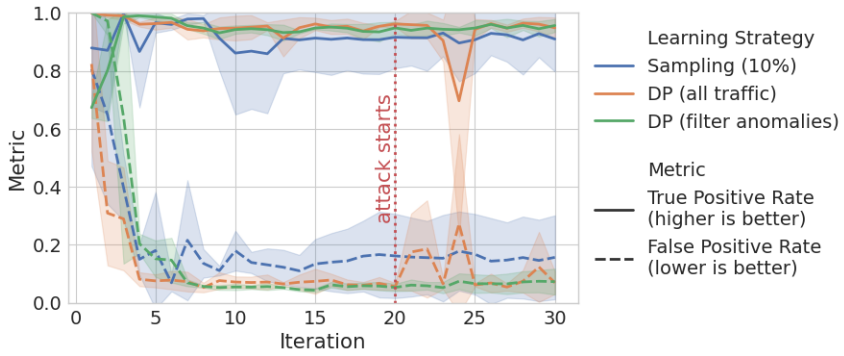
We split the search into two parts: *(i)* executing BO on the control plane to propose candidate distributions; and *(ii)* scoring the candidates in the data plane. Then, BO uses the scores to update the candidates and we repeat the process.

Use case: Anomaly detection Our approach allows detecting anomalies in the data plane. The key insight is that anomalies follow

[1] Measurable properties like ports, size, inter-packet time, flow duration, etc.



(a) Anomaly detection degrades significantly if the distribution is learned from data containing a high fraction of attack packets.



(b) We can efficiently learn a distribution from data-plane scores (DP). Combined with anomaly detection, we avoid poisoning with attack traffic. Sampling achieves worse performance because it misses a lot of traffic. Lines (shaded areas) show mean (std) over three runs.

a different distribution than benign traffic and thus score differently. We calculate a confidence interval on the score of expected traffic and classify any traffic scoring outside of this interval as anomalous. By deferring this until a couple of packets arrive, we can reduce false positives on single packets and only alert on anomalous sequences.

Further, we protect our learning mechanism from being poisoned by attack traffic: We constantly update the learned models to adapt to benign traffic shifts but discard updates from anomalous traffic.

Control-plane: Bayesian optimization Learning distributions via BO can be stated as $\text{argmax}_{x \in \mathcal{X}} f(Q_x)$, where $\mathcal{X}$ is the space of distribution parameters. From the perspective of BO, f is a noisy black-box objective function representing the quality of x. Q is user-defined and may range from a single distribution to a mixture of different distribution families, depending on the feature.

At each step, BO considers past measurements of $f(Q_x)$ to generate new sets of candidate parameters. For each candidate x, we evaluate f on the actual traffic in the data plane. The results are then used by BO to refine the parameters.

Data-plane: Evaluating distributions Given a set of candidates x, the data plane needs to measure the quality $f(Q_x)$ on the observed traffic. This faces two main challenges:

First, it is hard to calculate how well a proposed model Q_x matches an unknown traffic distribution P. To overcome that, we represent $f(Q_x)$ with a *proper scoring rule* [3]. Proper scoring rules can be computed from Q_x and observed features, and a model with a higher score is guaranteed to be closer to P, with a distance measure associated with the scoring rule. Specifically, we use the *logarithmic score*: $\tilde{f}(Q_x, y) = \frac{1}{|Y|} \sum_y \log Q_x(y)$, i.e., the average log probability of observed features (y) under the proposed model Q_x. This score is associated with the *Kullback-Leibler (KL) divergence*, i.e., the Q_x that maximizes the score has the minimal KL divergence to P.

Second, the scores may be computationally complex functions that cannot be easily evaluated. For example, programmable data planes are unable to compute logarithms. To bypass this issue, when the control plane decides on a set of evaluation points x, it precomputes the scores, in our case $\log Q_x(y)$, for various ys, converting the parametrized distribution into a log-probability lookup table. As a result, scoring the distribution is just a lookup and add, allowing for calculating scores in the data plane, on *all* traffic. We can even score multiple distributions in parallel for anomaly detection and BO. Each additional distribution allows BO to explore more areas of the parameter space at once.

3 PRELIMINARY EVALUATION

We evaluate the feasibility of our approach by simulating a traffic feature that follows a Zipf distribution with $a = 1.3$. In addition, we simulate attack traffic that follows another Zipf distribution with $a = 1.1$. This feature might represent destination ports per source: For benign sources, the distribution is narrow, as most send packets to widely used ports. For attack sources, e.g., a port scanning attacker, the distribution is naturally wider.

We score three distributions in the data plane: the current best estimate for anomaly detection, and two others for exploration.

Anomaly detection Our approach works well if we can learn an accurate distribution of benign traffic. The ROC curve in Figure 1a shows the true and false positive rate for confidence interval (CI) thresholds from 0–100%. After observing only 10 packets per flow and using a 95% CI, we detect 94% of anomalies with 5% false positives (green line, top left). Performance degrades with worse distributions, e.g., if we learn from attack traffic.

Learning over time Figure 1b shows that we can learn accurate distributions. We bootstrap learning with 20 iterations of 1000 benign packets each, followed by 10 iterations of 1000 benign and attack packets each. By combining scoring with anomaly detection, we avoid poisoning the model with attack traffic. Learning in the control plane with a sampling rate of 10% misses the distribution tail, degrading performance even before the attack starts.

4 CONCLUSION AND FUTURE WORK

We have shown that we can learn distributions to detect anomalies in the control plane via BO by leveraging the data plane to score distribution candidates on *all* traffic without sampling.

In future work, we want to explore the potential and limitations of our approach. How complex distributions can we learn with a given amount of data-plane resources for scoring? Furthermore, we assume that attacks can be identified by sharp changes in the distribution, such that we reject anomalies while learning benign shifts. Does this assumption hold in practice?

REFERENCES

[1] Albert Gran Alcoz, Martin Strohmeier, Vincent Lenders, and Laurent Vanbever. 2022. Aggregate-Based Congestion Control for Pulse-Wave DDoS Defense. In *Proceedings of the ACM SIGCOMM 2022 Conference (SIGCOMM '22)*. Association for Computing Machinery, New York, NY, USA, 693–706. https://doi.org/10.1145/3544216.3544263

[2] Gilberto Fernandes, Joel J. P. C. Rodrigues, Luiz Fernando Carvalho, Jalal F. Al-Muhtadi, and Mario Lemes Proença. 2019. A Comprehensive Survey on Network Anomaly Detection. *Telecommunication Systems* 70, 3 (March 2019), 447–489. https://doi.org/10.1007/s11235-018-0475-8

[3] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (March 2007), 359–378. https://doi.org/10.1198/016214506000001437

[4] Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. 2017. Detecting HTTP-based Application Layer DoS Attacks on Web Servers in the Presence of Sampling. *Computer Networks* 121 (July 2017), 25–36. https://doi.org/10.1016/j.comnet.2017.03.018