Ph.D. Thesis Position 2022

Title: Modeling Modern Network Traffic for Automated Machine Learning in Network Systems.

Host laboratory: LISTIC, Polytech Annecy-Chambéry, 5 Chemin de Bellevue, 74940 Annecy-le-Vieux

Advisor(s):
Francesco Bronzino, Assistant Professor
Université Savoie Mont Blanc, Annecy-le-Vieux `fbronzino@univ-smb.fr`

External Collaborators:
The thesis will be carried on within a bi-lateral ANR collaboration between Université Savoie Mont Blanc, University of Chicago, and Stanford University.

Description. Both network operations and research depend on the ability to answer questions about network traffic. Decades ago, the questions were simpler: they involved traffic volumes and simple performance metrics. The answers were also more apparent: most traffic was not encrypted, and the answers to most questions were readily apparent from protocol headers and unencrypted packet payloads. Today, operators and researchers are asking more sophisticated questions about application performance, quality of experience (QoE), and malicious traffic originating from IoT devices, as well as trying to predict the impact of potential changes. And yet, as questions are becoming increasingly complex and important, network data is becoming more difficult to obtain. Increased traffic requires operators to make hard decisions about sampling and altogether precludes analyzing individual packets and reassembled streams. Furthermore, traffic is increasingly opaque. Web content has become ubiquitously encrypted, preventing operators from directly inspecting video streams to troubleshoot performance problems. Major services have moved to a handful of IP addresses on large cloud providers like Amazon, Google, and Cloudflare, removing the identity once provided by IP addresses. Networks contain increasingly heterogeneous manufacturer-controlled devices that cannot be troubleshooted locally. As a result, even seemingly simple, but important questions like "What content is sent in cleartext?" or "What is the packet loss for Netflix traffic on my network?" are impossible to answer today.

Despite traffic becoming more opaque, it is possible to infer many of the characteristics of traffic most important to operators through statistical learning. Consider for example, monitoring video steaming quality. This has become increasingly difficult as the recent adoption of HTTPS and QUIC prevents directly observing video quality metrics. Our recent work shows that it is possible to infer startup delay and resolution of encrypted video to Netflix and Youtube in real-world homes by training a model on traffic features [2]. And yet, this model and much other previous work on applying machine learning to network operations and security has not made the transition to practice at ISPs. Most models do not perform outside of the isolated laboratory environment in which they were trained and even when robust, they require access to data that cannot be collected and analyzed in real-time on high-speed networks. Building and operationalizing inference models is more than a "simple matter of engineering". Some of these challenges include: (1) evaluating models at high speeds, including perform- ing flow reassembly at high speeds; (2) coping with dirty data, such as network traces that include other network traffic or training data that is unlabeled or (worse) erroneously labeled; (3) representing network traffic data in formats that are amenable to training; and (4) determining when to retrain these models, due to drift in network traffic patterns over time.

In this Ph.D. thesis, the student will develop new methods that aim to make it easier for operators and researchers to ask questions about network traffic. Doing so involves solving new, challenging research questions to create the required building blocks to model and process traffic on modern networks. Towards this, this thesis will focus on studying new methods for deploying models in operational networks. We will use the software platforms and algorithmic primitives we built in our previous work [1] to design new techniques

and tools for operators to solve the challenges that block them from transferring developed models from isolated laboratory experiments to real-world deployments. We will support their need to monitor their networks and investigate problems in real time by: (1) extending automated model selection to account for systems costs and real world limitations; (2) addressing the need to be able to determine when models become inaccurate and distinguishing model inaccuracies from problems that are inherent to the network; (3) improve models robustness by investigating a generalized approach for model transfer.

Candidate Requirements.

- The candidate should have completed a qualifying program by the starting date of the thesis.
- Comfortable speaking English or French (French is not required).
- Good understanding of at least one between computer networks protocols and systems or machine learning methods (preferably both)
- Good proficiency with at least one programming language, preferably Golang or Rust.

What to submit. An up to date CV, university transcripts, and a letter of motivation clearly stating what the motivations to work on the described subject.

References

[1] F. Bronzino, P. Schmitt, S. Ayoubi, H. Kim, R. Teixeira, and N. Feamster. Traffic refinery: Cost-aware traffic representation for machine learning in networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(3), Dec. 2021.

[2] F. Bronzino, P. Schmitt, S. Ayoubi, G. Martins, R. Teixeira, and N. Feamster. Inferring streaming video quality from encrypted traffic: Practical models and deployment experience. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3), Dec. 2019.