PINOT: Programmable Infrastructure for Networking

https://pinot.cs.ucsb.edu

Roman Beltiukov
UC Santa Barbara
USA

Sanjay Chandrasekaran
UC Santa Barbara
USA

Arpit Gupta
UC Santa Barbara
USA

Walter Willinger
NIKSUN Inc.
USA

ACM Reference Format:
Roman Beltiukov, Sanjay Chandrasekaran, Arpit Gupta, and Walter Willinger. 2023. PINOT: Programmable Infrastructure for Networking: https://pinot.cs.ucsb.edu. In *Applied Networking Research Workshop (ANRW '23), July 22–28, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3606464.3606485

As modern network communication moves closer to being fully encrypted and hence less exposed to passive monitoring, traditional network measurements that rely on unencrypted fields in captured traffic provide less and less visibility into today's network traffic. At the same time, approaches that use techniques from machine learning (ML) to extract subtle temporal and spatial patterns from encrypted packet-level traces have shown great promise in offsetting the lack of visibility due to encryption [1–3, 5–7, 10–15, 18, 23, 24].

Despite their promise, ML-based approaches often have a credibility problem that arises from the quality of underlying training data. Given the challenges of curating high-quality training data at scale, researchers typically end up collecting their own (or reusing existing third-party or synthetic) data, often from small-scale testbeds. Such data is generally of low quality as it is not representative of the target environment, collected over too short of a time period, or measured at too coarse of a granularity. The learning models trained using such data tend to be vulnerable to different failure modes that make them not credible [8]. This observation begs a fundamental question, **how can we develop credible ML artifacts for managing encrypted network traffic?**

This paper describes our ongoing efforts to enable researchers and practitioners to develop more credible ML artifacts by lowering the effort that is required for collecting more high-quality data for a wide range of learning problems from realistic and representative network environments.

ANRW '23, July 22–28, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0274-7/23/07.
https://doi.org/10.1145/3606464.3606485

1 PINOT

To answer the stated question, we argue for developing a representative open infrastructure where researchers can freely explore different data collection strategies (e.g., collecting the same type of data as some existing data from new networks) or flexibly perform new data collection experiments (i.e., reproducing existing experiments) to develop more credible ML-based solutions. Previous efforts [9] consider university campus networks as representative real-world infrastructures, mainly because of their scale, user base, and realistic nature of observed traffic.

At UCSB, we designed and implemented PINOT[1], a programmable data-collection infrastructure for collecting labeled data from the underlying production network. Being able to collect and curate such data sets aids the development of more credible ML models for different networking-related learning problems. For example, we have used this platform to curate labeled data sets to train learning models for inferring QoE metrics for various video streaming (e.g., YouTube, Twitch, etc.) and video conferencing applications (e.g., Google Meet); detecting traffic of interest for cybersecurity appliances (e.g., IDS); and performing device or application fingerprinting. We also used a combination of active and passive measurements to first find campus locations where devices experience large downstream RTT values and then identify a device, access point, or downlink as the main culprit for the observed large RTT. Being highly configurable, the platform allows us, for example, to collect data over extended periods of time, enabling the study of weekly, monthly, and seasonal trends or changes in traffic patterns.

2 PASSIVE DATA COLLECTION

To enable passive real-world traffic monitoring, PINOT supports data collection from the border gateway of the UCSB campus network. This data collection infrastructure relies primarily on PISA-based switches to ensure the passive collection of packet-level traffic traces in a privacy-preserving manner at scale. More concretely, for each incoming packet, the switch generates a cloned copy of the packet with randomized privacy-sensitive fields (e.g., IP addresses of end users on campus) and pruned payload fields, and then load-balances the output packet stream to a cluster of collection servers.

[1]Supported by the NSF's Campus Cyberinfrastructure (CC*) grant, OAC-2126327

Each of these servers captures and stores anonymized packet headers for further analysis. We augment these packet traces by joining them with logs from various production appliances (e.g., Aruba's AirMon, and Palo Alto's MTA) to attach labels for learning problems.

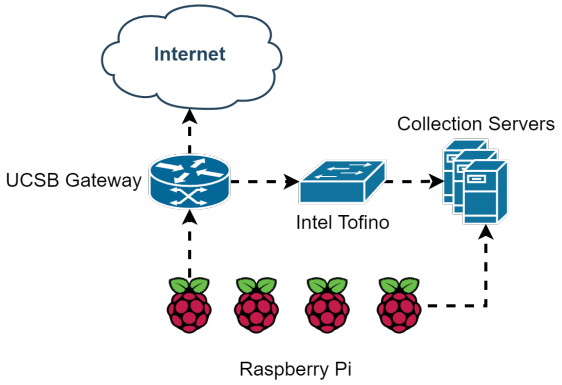

Figure 1: Simplified view of the PINOT infrastructure.

3 ACTIVE DATA COLLECTION

To support active measurements, we deployed a fleet of network measurement devices across the campus and connected them to the public network infrastructure.

We used ARM64-based single-board computers Raspberry Pi 4, running a modified version of the Ubuntu Server that includes performance and security optimizations. Each device is powered either via Power-over-Ethernet or an outlet adapter to flexibly choose deployment locations across the campus. The choice of devices and OS enables us to run arbitrary network experiments, from simple ones such as pings to more complex scenarios such as synchronized simulations of network attacks.

To control these devices, we deployed a SaltStack configuration management system [22] with agents installed on our nodes. This system allows us to control nodes remotely, automatically deploy and start network experiments, and collect produced artifacts remotely. SaltStack's agent-based approach is also particularly well suited for NAT-based network environments so that devices can be deployed without publicly available addresses. We also use our `netunicorn` platform [17] to implement network experiments and seamlessly switch between physical infrastructures and virtual alternatives (e.g., Containernet [19] or clouds).

We deployed these devices across UCSB's campus where the university operates a campus-wide Wi-Fi infrastructure for all students, available via access points in each building and facility of the university. During the deployment, we connect our nodes to both wired access ports and the public Wi-Fi network, sharing the network connection with all university students. This allows us to mimic a typical network user in the campus network and obtain instances of representative real-world measurements from different physical locations and under a range of different network conditions.

During the first phase of our deployment, we identified the most populated and traffic-heavy areas within the campus (e.g., student dormitories, libraries, university centers, and dining halls). In total, we deployed over 50 devices in these locations that are now online and allow us to perform various measurement experiments and tests. We also provided each device with a QR code that links to a device portal where we provide different statistics and results of background measurements to other users at this location.

4 DISCUSSION & FUTURE RESEARCH

To further democratize ML-based networking research [9], we made the PINOT platform publicly available. We offer access to the platform based on principles of fair and responsible usage. Compared to existing alternatives [4, 16], we support arbitrary experiments and allow to implement them with `netunicorn` platform.

Users are able to leverage PINOT's active and passive data collection capabilities separately or in combination. While passively collected data has limitations (i.e., lack of payload and sensitive headers), it could be used for enriching labeled datasets with unlabeled user traffic or comparing captured traffic from multiple vantage points to investigate different latency issues or network anomalies.

At the same time, our use of an existing campus network infrastructure constrains the type of data collection scenarios that PINOT can support. For example, we cannot support specific data collection scenarios such as data-center-oriented traffic collections or efforts that presume a particular type of network connectivity.

We encourage and make it possible for all users to participate in the platform development, either by receiving a physical node or by hosting a virtual node using our provided Docker container. Virtual nodes entail only minimal configuration efforts, require no special permissions, and only assume basic Internet access.

We maintain a project website [21] where we share information about the project, including details of our current installation at UCSB and up-to-date statistics and meta-data webpages for each deployed device. For network experimenters and developers, we also share the OpenAPI endpoint [20] for our raw network data and detailed information on node location for more controlled and automated node selection for measurements.

Our ongoing and future research efforts focus on increasing the coverage within our campus network. We also plan to improve the flexibility of our network setup to provide more capabilities for potential network experiments by adding options to redirect the required traffic from our nodes through a single controlled routing point to implement different traffic-shaping solutions or emulate a range of network conditions and scenarios.

REFERENCES

[1] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé. Mobile encrypted traffic classification using deep learning. In *2018 Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–8, 2018.

[2] R. Amin, E. Rojas, A. Aqdus, S. Ramzan, D. Casillas-Perez, and J. M. Arco. A survey on machine learning techniques for routing optimization in sdn. *IEEE Access*, 9:104582–104611, 2021.

[3] B. Anderson and D. McGrew. Machine learning for encrypted malware traffic classification: Accounting for noisy labels and non-stationarity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1723–1732, New York, NY, USA, 2017. Association for Computing Machinery.

[4] Ripe atlas. https://atlas.ripe.net/.

[5] 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. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(3), dec 2019.

[6] Z. Bu, B. Zhou, P. Cheng, K. Zhang, and Z.-H. Ling. Encrypted network traffic classification using deep and parallel network-in-network models. *IEEE Access*, 8:132950–132959, 2020.

[7] Z. Chen, K. He, J. Li, and Y. Geng. Seq2img: A sequence-to-image based approach towards ip traffic classification using convolutional neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1271–1276, 2017.

[8] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning, 2020.

[9] A. Gupta, C. Mac-Stoker, and W. Willinger. An effort to democratize networking research in the era of ai/ml. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, HotNets '19, page 93–100, New York, NY, USA, 2019. Association for Computing Machinery.

[10] J. Kim, Y. Jung, H. Yeo, J. Ye, and D. Han. Neural-enhanced live streaming: Improving live video ingest via online learning. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '20, page 107–125, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Y. Li, H. Liu, W. Yang, D. Hu, and W. Xu. Inter-data-center network traffic prediction with elephant flows. In *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, pages 206–213, 2016.

[12] C. Liu, L. He, G. Xiong, Z. Cao, and Z. Li. Fs-net: A flow sequence network for encrypted traffic classification. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 1171–1179, 2019.

[13] A. Mahimkar, A. Sivakumar, Z. Ge, S. Pathak, and K. Biswas. Auric: Using data-driven recommendation to automatically generate cellular configuration. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, SIGCOMM '21, page 807–820, New York, NY, USA, 2021. Association for Computing Machinery.

[14] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, page 197–210, New York, NY, USA, 2017. Association for Computing Machinery.

[15] B. Miller, L. Huang, A. D. Joseph, and J. D. Tygar. I know why you went to the clinic: Risks and realization of https traffic analysis. In E. De Cristofaro and S. J. Murdoch, editors, *Privacy Enhancing Technologies*, pages 143–163, Cham, 2014. Springer International Publishing.

[16] Netrics. https://github.com/chicago-cdac/nm-exp-active-netrics.

[17] netunicorn. https://github.com/netunicorn.

[18] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, and K. Wehrle. Website fingerprinting at internet scale. In *NDSS*, 2016.

[19] M. Peuster, H. Karl, and S. van Rossem. MeDICINE: Rapid prototyping of production-ready network services in multi-PoP environments. In *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pages 148–153, Nov. 2016.

[20] Pinot swagger. https://pinot.cs.ucsb.edu/api/v1/info/swagger/#/.

[21] Pinot project website. https://pinot.cs.ucsb.edu.

[22] Salt project. https://saltproject.io/.

[23] Y. Shi and S. Biswas. Protocol-independent identification of encrypted video traffic sources using traffic analysis. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6, 2016.

[24] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin. Intrusion detection by machine learning: A review. *Expert Systems with Applications*, 36(10):11994–12000, 2009.