Poster: Understanding and Managing Changes in IoT Device Behaviors for Reliable Network Traffic Inference

Authors: Shayan Azizi, Norihiro Okui, Masataka Nakahara, Ayumu Kubota, Gustavo Batista, Hassan Habibi Gharakheili

Contact: Shayan Azizi (*s.azizi@unsw.edu.au*)

- Data-driven inference from network traffic is increasingly becoming a practical, cost-effective, and scalable method for various networking use cases, such as IoT device fingerprinting.
- Machine learning (ML) models are often trained on a batch of data in one environment, and then used to make predictions on ***future*** data of the same environment or ***other environments***.
  - *machine learning models often underperform in changing conditions.*
- The reason is believed to be changes in the underlying distribution of network traffic data, also known as ***Concept Drifts***.

Image by pngtree.com

No prior research has thoroughly studied concept drifts in IoT network traffic data.

[RQ1] *What are the effective ways to quantify and characterize concept drifts in IoT network traffic data?*

How to capture the distribution of network traffic data?

How to measure distributional discrepancies?

Curse of dimensionality

How to tailor quantification and characterization of concept drifts to facilitate managing their adverse effects?

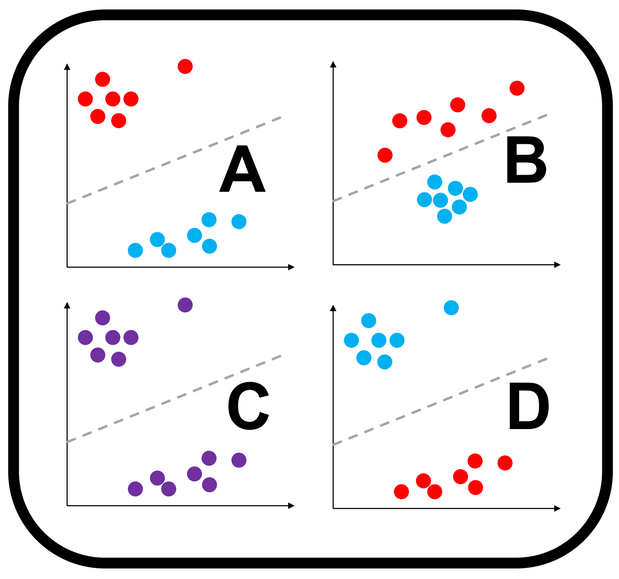
After changes in the network traffic data are effectively quantified and characterized, one might ask …

[RQ2] *Whether, when, and how different ML models are affected by concept drifts in IoT network traffic data?*

Question: How should one handle concept drifts?
***Naïve* answer:** Models should be updated using fresh data regularly.
Issues: (1) Labeled data is often scarce; (2) Regular re-training can be sub-optimal and inefficient.
Conclusion: network operators need to know when updating the inference models is essential.

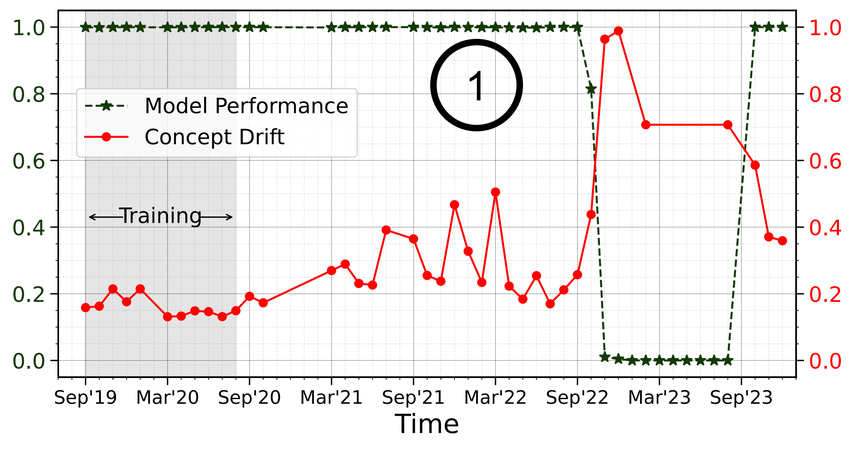
Challenge: With limited access to ground-truth data, determining the optimal time to refresh models is nontrivial.

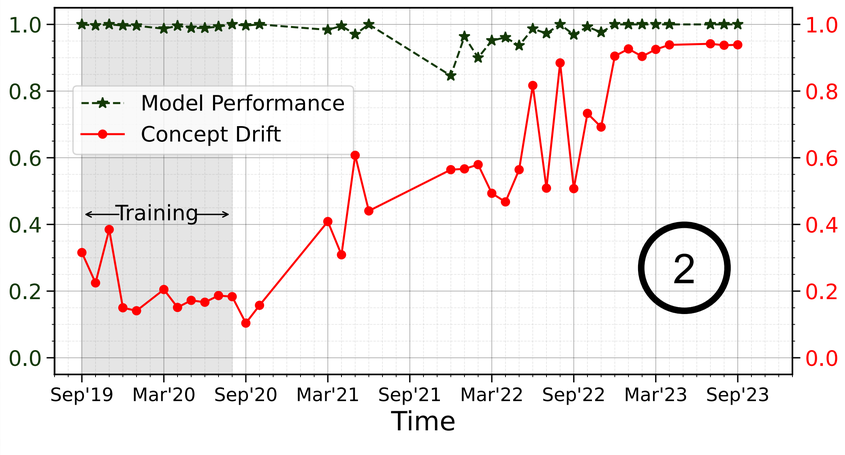
[RQ3] *How can we effectively signal for and implement model adaptation with minimum reliance on labeled data?*

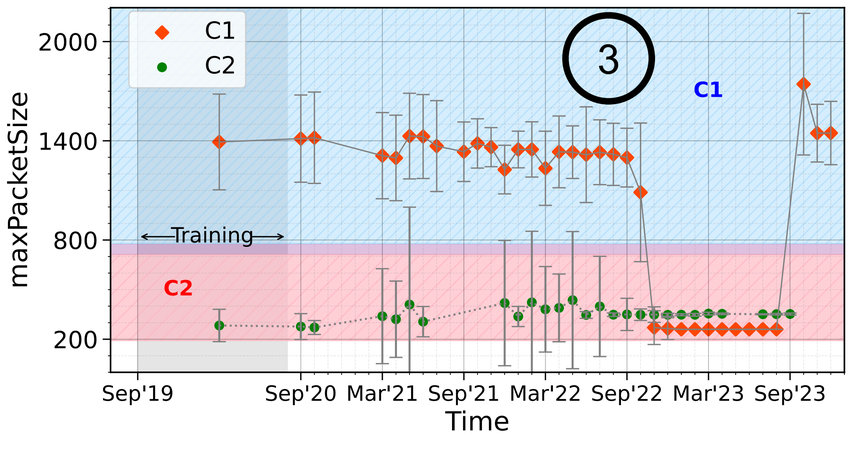


To answer these questions, we have access to an IoT testbed dataset compiled over **five years**, along with 25 real smart home datasets.

Our data consists of YAF-exported **IPFIX** flow records, with IP and MAC address information obfuscated to simulate a **post-NAT** telemetry scenario.







IE	Name
1	packetTotalCount
2	octetTotalCount
3	smallPacketCount
4	largePacketCount
5	nonEmptyPacketCount
6	dataByteCount
7	averageInterarrivalTime
8	firstNonEmptyPacketSize
9	maxPacketSize
10	standardDeviationPayloadLength
11	standardDeviationInterarrivalTime