



UNSW  
SYDNEY



UNIVERSITY of CALIFORNIA  
IRVINE

# Congested by the Past: The Dataset Lag in Network Traffic Analysis

Nimesha Wickramasinghe<sup>†</sup>, Sanjay Jha<sup>†</sup>, Gene Tsudik<sup>‡</sup>, Arash Shaghaghi<sup>†</sup>

*{n.wickramasinghe, sanjay.jha, a.shaghaghi}@unsw.edu.au, gtsudik@ics.uci.edu*

<sup>†</sup> The University of New South Wales (UNSW), Sydney, Australia

<sup>‡</sup> University of California Irvine, USA



UNSW  
SYDNEY



NTC-IQ

<https://ntciq.web.cse.unsw.edu.au>

## NTC-IQ

### **Bridging the Gap Between Network Intelligence and AI**

NTC-IQ is an academic initiative aimed at enabling network-aware artificial intelligence. We present a research-driven framework to analyze and classify encrypted network traffic using RFC-compliant features - ensuring privacy preservation without sacrificing performance.

# Introduction

- Network traffic analysis (NTA) supports security, intrusion detection, and optimization.
- Machine learning drives modern NTA.
- However, datasets define the limits of what ML models can learn.

# Problem Statement

- Many widely used datasets (e.g., ISCX-VPN, ISCX-Tor, USTC-TFC) are pre-2018.
- They miss key protocol updates, such as,
  - TLS 1.3 [RFC 8446] - August 2018
  - QUIC [RFC 9000] - May 2021
  - HTTP/3 [RFC 9114] - June 2022
- Models trained and tested on these datasets may,
  - Fail to generalize to modern encrypted traffic
  - Embed outdated assumptions
  - Report inflated accuracies that do not hold in real-world conditions

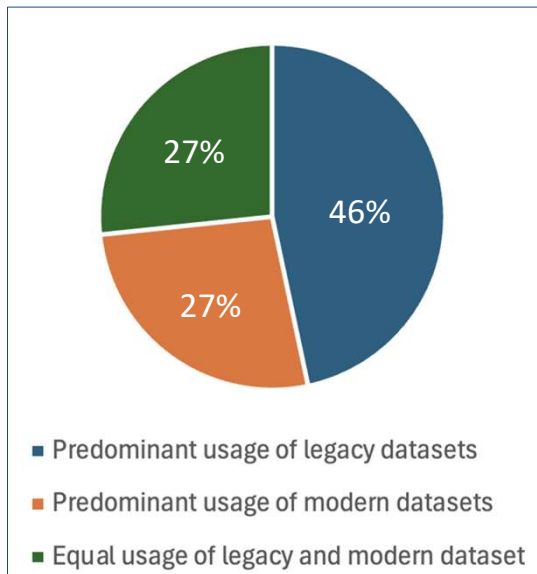
# Study Objective

- Examine how recent research (2024–2025) uses datasets.
- Identify the gap between modern traffic and academic benchmarks.
- Propose a call for action and research agenda for updated datasets.

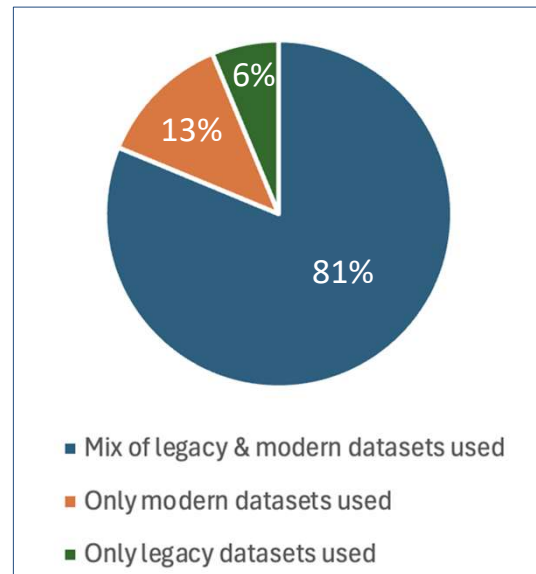
# Methodology

- Reviewed papers from top-tier venues:
  - Security: IEEE S&P, USENIX, CCS, NDSS
  - Networking: NSDI, SIGCOMM
  - Web: WWW
  - AI: AAAI
- Inclusion criteria:
  - “Network” or “traffic” in title
  - Focused on NTA (not neural networks, etc.)

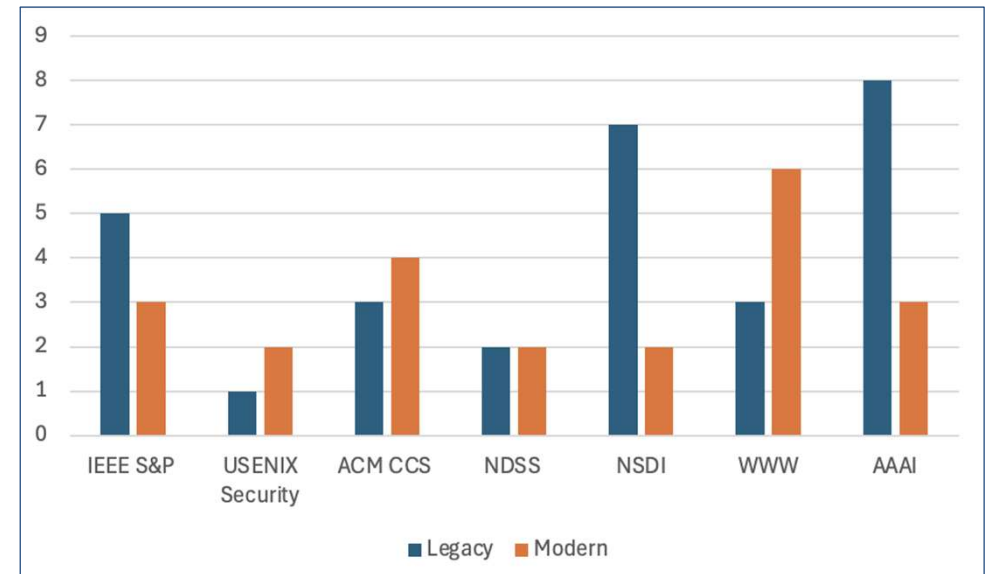
# Findings



Dataset usage dominance



Dataset usage trends



Dataset usage and venues

# Research Agenda

- Reevaluate classifiers on modern datasets:
  - AppSniffer - WWW'23 - VPN traffic
  - Oscar Multitab - CCS'24 - Tor traffic
  - HyperVision - NDSS'23 - Malware traffic
  - CipherSpectrum - IEEE SP'25 - Web traffic
- Treat datasets as living artifacts: update, retire, and share regularly.
- Encourage reproducibility and openness.

# Discussion & Call to Action

- Research remains anchored in outdated datasets, creating a gap between academia and real-world.
- Dataset selection must become a first-class concern and outdated benchmarks should be questioned.
- Community and industry collaboration are essential for sharing modern, privacy-conscious datasets.

# Project NTC-IQ

- An initiative that aims to bridge the gap between AI and real-world network intelligence.
- **IEEE SP'25 (Oakland)**
  - ***SoK: Decoding the Enigma of Encrypted Network Traffic Classifiers***
    - Dataset evaluation for validity
    - Empirical critique of ML based NTC design oversights
    - Recommend best practices
- **WWW'25**
  - ***Less is More: Simplifying Network Traffic Classification Leveraging RFCs***
    - Protocol-driven design grounded in RFC specifications
    - Address key misconceptions in current approaches

# CipherSpectrum Dataset

Most publicly available NTC datasets:

- Contain unencrypted or outdated traffic
- Include traffic sessions encrypted with deprecated cipher suites (e.g., RC4, 3DES)
- Suffer from class imbalances or limited raw traffic

CipherSpectrum addresses these issues by providing equal representation for each class, using TLS 1.3 traffic encrypted with AEAD cipher suites.

40 total classes

3000 sessions per class

120000 total sessions

**Access Dataset**





Nimesha Wickramasinghe



Sanjay Jha



Gene Tsudik



Arash Shaghaghi

- 
- Nimesha Wickramasinghe: [n.wickramasinghe@unsw.edu.au](mailto:n.wickramasinghe@unsw.edu.au)
  - Sanjay Jha: [sanjay.jha@unsw.edu.au](mailto:sanjay.jha@unsw.edu.au)
  - Gene Tsudik: [gtsudik@ics.uci.edu](mailto:gtsudik@ics.uci.edu)
  - Arash Shaghaghi: [a.shaghaghi@unsw.edu.au](mailto:a.shaghaghi@unsw.edu.au)



NTC-IQ

<https://ntciq.web.cse.unsw.edu.au>