The DARPA SEARCHLIGHT Dataset of Application Network Traffic

Calvin Ardi[1], Connor Aubry[2], Brian Kocoloski[1], David DeAngelis[1],
Alefiya Hussain[1], Matt Troglia[2], Stephen Schwab[1]

[1] University of Southern California/Information Sciences Institute
[2] Sandia National Laboratories

Workshop on Cyber Security Experimentation and Test (CSET'22) – August 8, 2022

USC Viterbi School of Engineering
Information Sciences Institute

Sandia National Laboratories

“There is No AI Without Data”[1]

- AI/ML needs high quality data for training
- networking has some data, but the data is often:
  - unlabeled
  - heavily anonymized
  - highly specific or contrived measurements
  - unavailable (proprietary, PII/IP)

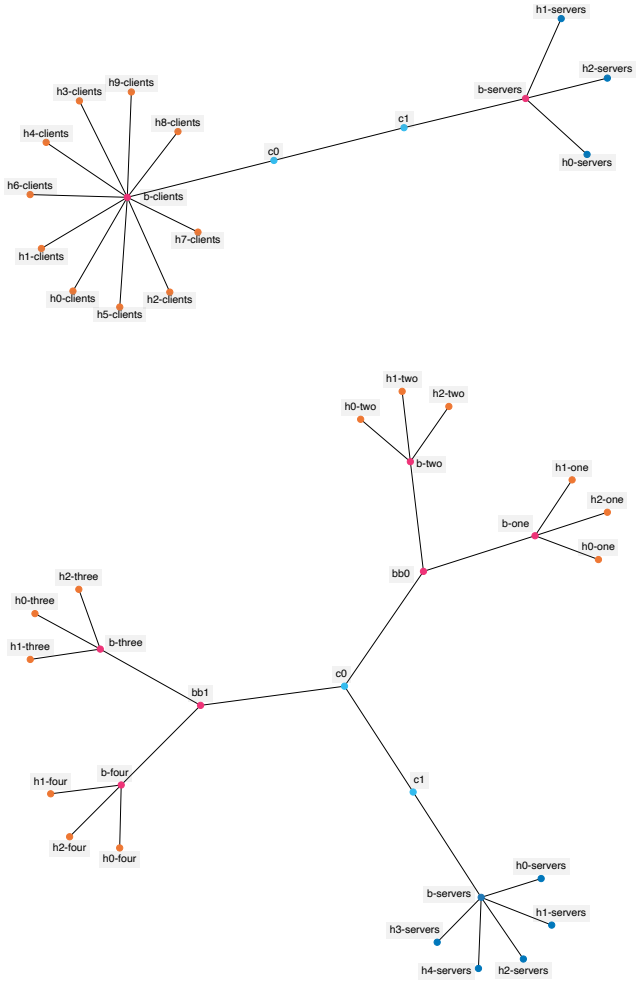
→ networking needs *high quality* and *usable* datasets for evaluating and comparing new algorithms and technologies

[1] Christoph Gröger. 2021. There is no AI without data. Commun. ACM 64, 11 (November 2021), 98–108. https://doi.org/10.1145/3448247

the DARPA SEARCHLIGHT dataset

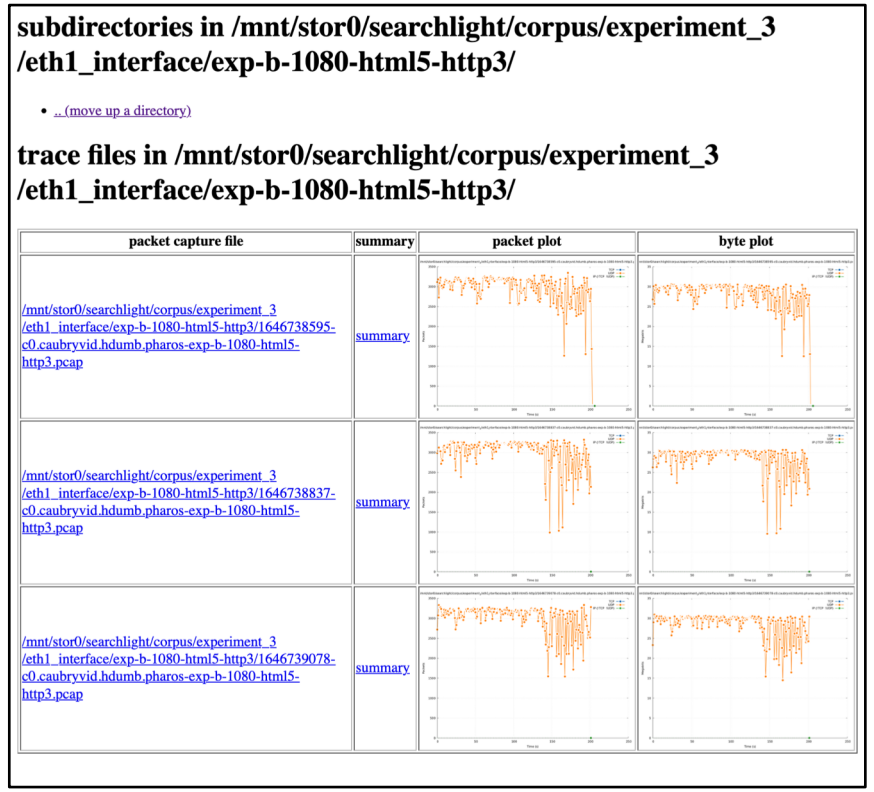
- generated on a network emulation testbed
- features:
  - *complete* – raw traffic from all sources/destinations
  - *labeled* – all flows are identified
  - varying levels of *complexity* – 132 scenarios over 2 network topologies, 3 apps (cloud doc. editing, video streaming, VTC), 2 VPNs
- use cases: AI/ML in traffic analysis and classification of plaintext/encrypted flows, network topology inference and path discovery, dynamic QoS enforcement

→ details in the paper!



accessing the DARPA SEARCHLIGHT dataset

- ~750 GB with ~2000 experiment runs of 132 scenarios
- multiple formats: .pcap, CSV, Parquet
- freely and publicly accessible at:
  - http://mergetb.org/projects/searchlight
  - AWS Open Data (soon)
- how can we help with enabling your networking research?



subdirectories in /mnt/stor0/searchlight/corpus/experiment_3 /eth1_interface/exp-b-1080-html5-http3/

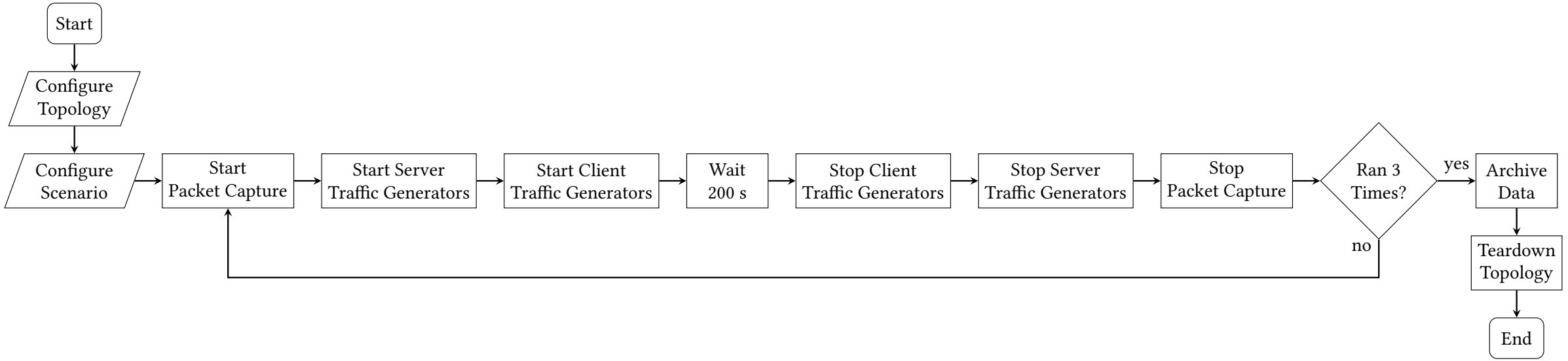
- .. (move up a directory)

trace files in /mnt/stor0/searchlight/corpus/experiment_3 /eth1_interface/exp-b-1080-html5-http3/

packet capture file	summary	packet plot	byte plot
/mnt/stor0/searchlight/corpus/experiment_3 /eth1_interface/exp-b-1080-html5-http3/1646738595-c0.caubryvid.hdumb.pharos-exp-b-1080-html5-http3.pcap	summary		
/mnt/stor0/searchlight/corpus/experiment_3 /eth1_interface/exp-b-1080-html5-http3/1646738837-c0.caubryvid.hdumb.pharos-exp-b-1080-html5-http3.pcap	summary		
/mnt/stor0/searchlight/corpus/experiment_3 /eth1_interface/exp-b-1080-html5-http3/1646739078-c0.caubryvid.hdumb.pharos-exp-b-1080-html5-http3.pcap	summary		

additional material

scenario execution flowchart



network traffic applications and configuration

Table 3: Applications and Configurable Features

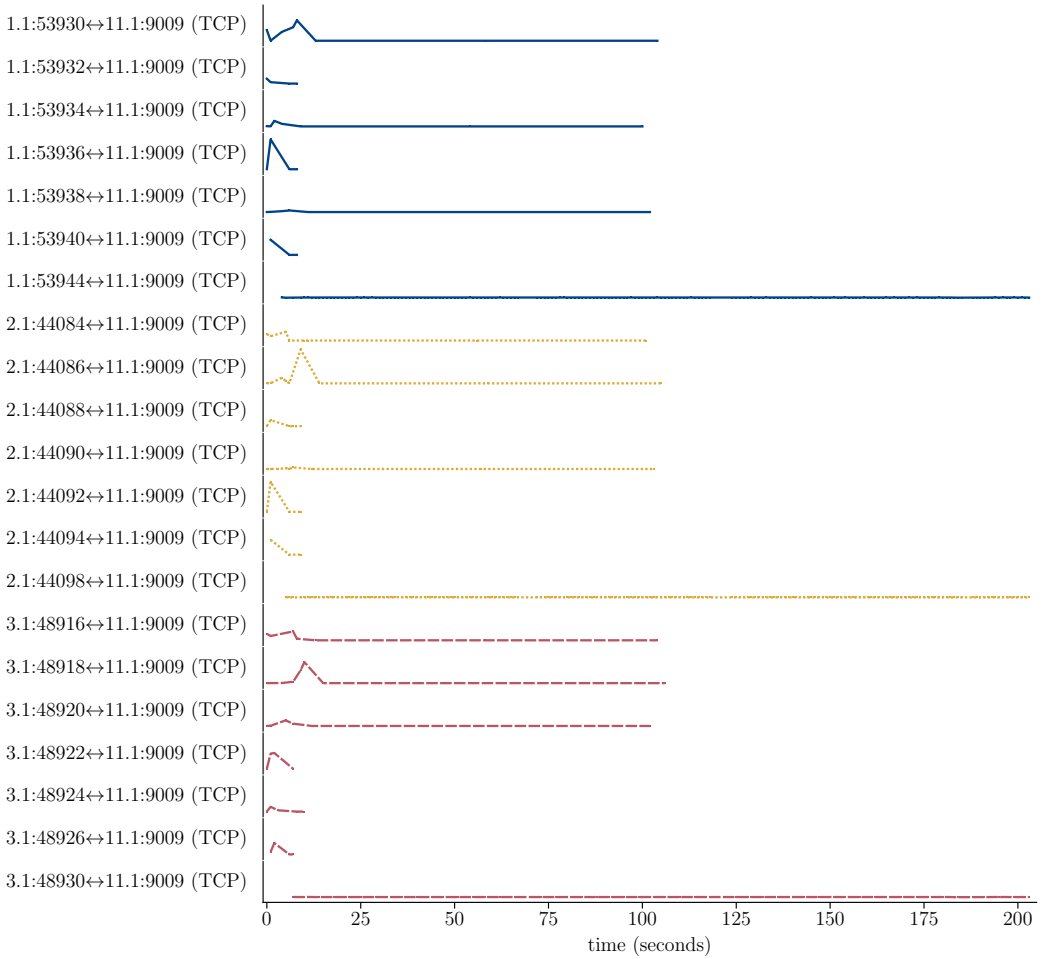
App.	Configurable Settings
`cbs`	speed := { fast (400 char/min), medium (200), slow (60) } bursty := { true, false } length := { *n* chars }
`vtc`	video := { true, false }
`video`	resolution := { 576p, 720p, 1080p } streaming := { DASH, HLS, HTML5 } transport := { HTTP/1.1, HTTP/1.1+TLS, HTTP/2, HTTP/3 }

Table 4: Number of Clients per Experiment Type

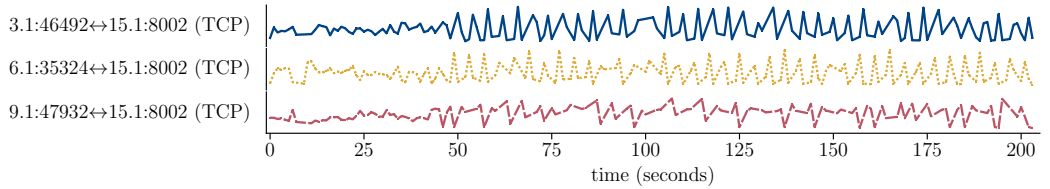
App.	exp-a	exp-b	exp-c
`cbs`	1	3	10
`vtc`	3	5	8
`video`	1	3	5

cbs – cloud-based document editing
vtc – video teleconferencing
video – video streaming

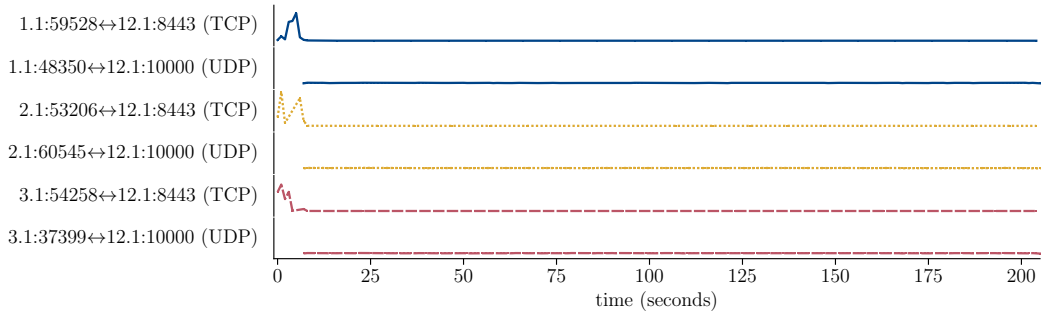
visual representation of experiment flows



cloud-based document editing (cbs)

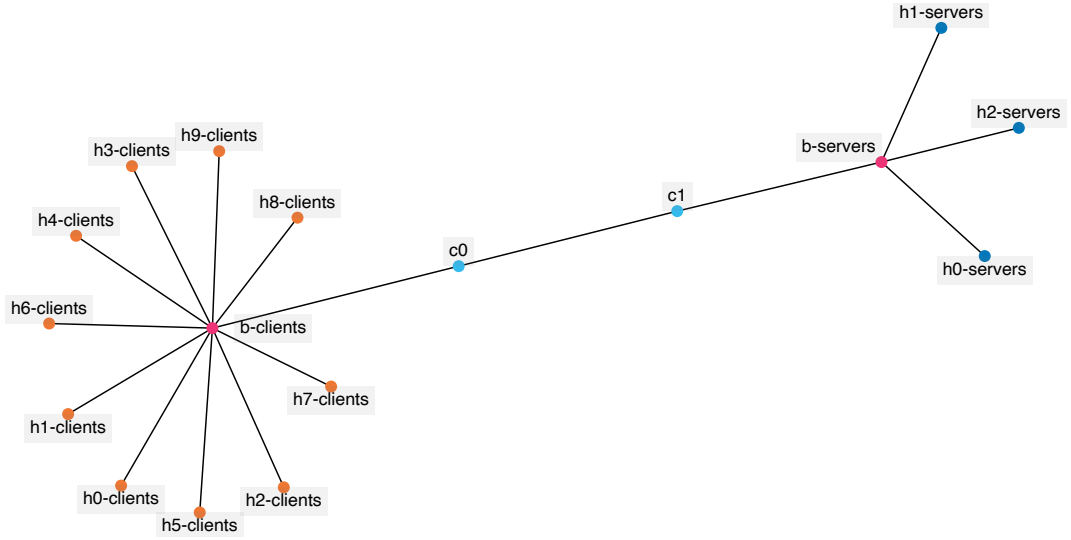


video streaming (video)

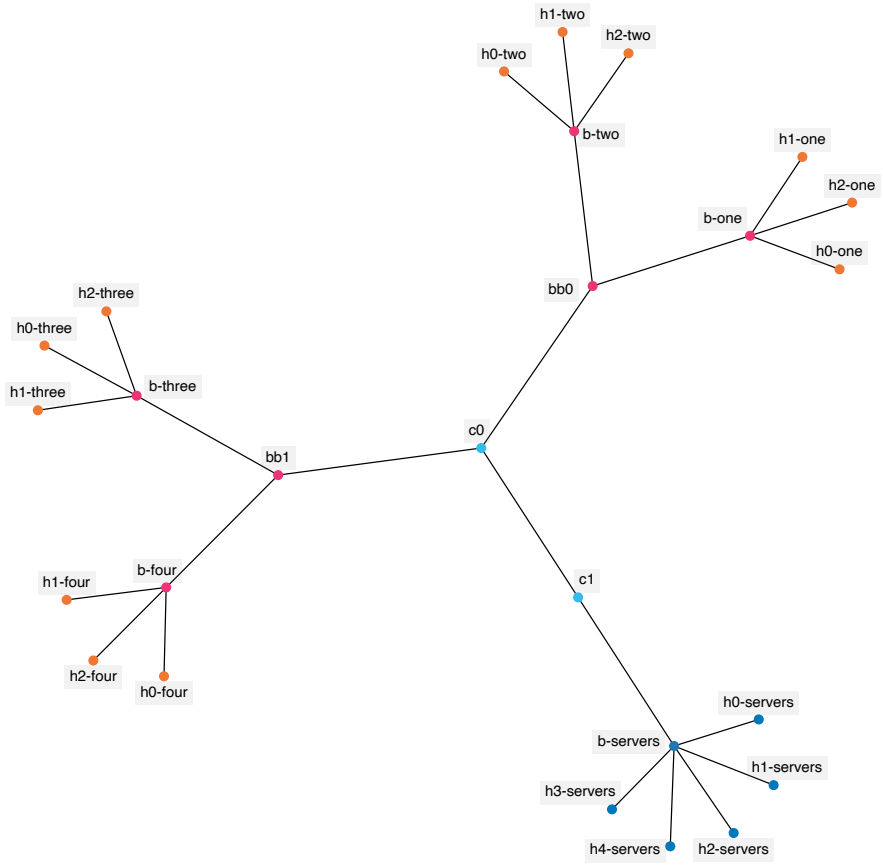


video teleconferencing (vtc)

network topologies



heavy dumbbell



tiered