Fall 11-1-2023

General Packet Datasets for Network Research

Bruce Hartpence
Rochester Institute of Technology, bhhics@rit.edu

Daryl Johnson
Rochester Institute of Technology, daryl.johnson@rit.edu

Bill Stackpole
Rochester Institute of Technology, wrsics@rit.edu

Recommended Citation
10.57673/gccis-yg55 OR www.doi.org/10.57673/gccis-yg55

This document describes the content of the network traffic datasets included in this collection and the conditions under which the packets were collected. These datasets were assembled from 2020 onward and so represent contemporary network traffic. There are periodic updates or additions to the dataset collection.

This collection contains curated (ex. enforced balance) and non-curated (raw) datasets. Both text and pcap (pcapng) file types can be opened with Wireshark.

When referencing these datasets, please use the following DOI:

10.57673/gccis-yg55 OR www.doi.org/10.57673/gccis-yg55

Contact info: Bruce Hartpence bhhics@rit.edu

Daryl Johnson dgjics@rit.edu

Bill Stackpole wrsics@rit.edu

Topology for Datasets 0-6

Datasets 0-6 were created via packet capture on a local RIT network testbed. The testbed consisted of Cisco routers, Cisco switches, computers and virtual machines. The primary operating system was Linux.

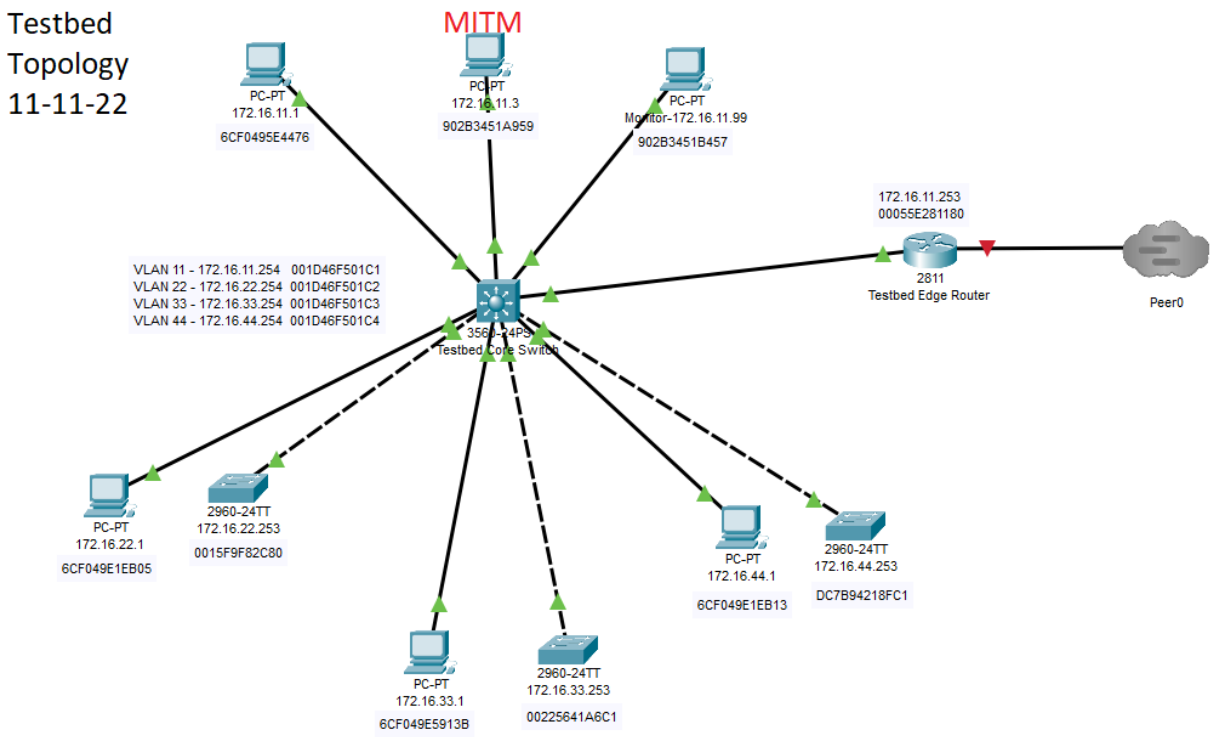

Figure 1 – Topology for Datasets 0-6

Dataset 0

This dataset used for training machine learning models. It is a curated, balanced dataset consisting of 5,000 samples of 14 different packet types or classes for a total of 70,000 packets.

Class labels:	Spanning Tree Protocol	Cisco Discovery Protocol
	Address Resolution Protocol	Loopback
	Internet Group Management Protocol	Internet Control Message Protocol (Type 0, 8)
UDP-based	Simple Service Discovery Protocol	NetBIOS Name Service
	Dynamic Host Configuration Protocol	Domain Name Service
TCP-based	HTTP/HTTPS/TCP data on ports 80, 8080 and 443	

Dataset 1

This is the curated, balanced validation set for machine learning models. It's construction is the same as Dataset 0 though smaller, having 500 samples for each of the 14 classes totaling 7000 packets.

Note: The timestamps for Datasets 0,1 are not sequential as the dataset was assembled from capture collections. Datasets 2-6 have fewer modifications although protocols outside of the desired classes were removed. Some of the datasets have protocol additions for testing. Timestamps for Datasets 2-6 are often sequential but have some variation where modification were made. Later datasets are raw captures.

Dataset 2

This is a larger dataset consisting of 99,998 packets containing the same classes noted previously. This is the first of the unbalanced datasets. For example there are 406 ARP packets, 3169 UDP based packets and 95,311 TCP based packets.

Dataset 3

This unbalanced dataset is similar in construction to Dataset 2 being weighted towards TCP based packets. It contains 25,847 packets.

Dataset 4

This unbalanced dataset contains 28,145 packets and is weighted toward TCP based traffic.

Dataset 5

This unbalanced dataset contains 24,354 packets but does not contain TCP based traffic. Instead, this dataset contains 14,534 Spanning Tree Protocol frames and 4,073 UDP based packets along with a collection of the classes noted earlier.

Dataset 6

This unbalanced dataset contains 18,610 packets and is weighted towards TCP based traffic. It does not contain 802.3 traffic.

capture1-10-11-23.pcapng

This dataset is a large collection of packets captured in the networking lab at RIT. It is a raw, non-curated capture of network traffic from Windows and Linux nodes consisting of 978,827 packets. This dataset is NOT restricted to the classes noted above.

Nodes in this topology sit behind a NAT box but connect to the Internet and each other. Several virtual machines were started and added to the topology. Activities on the topology include access youtube.com, PINGing (ICMP echo) between nodes, startup and shutdown.

The topology for capture1-10-11-23 is shown below. To facilitate a single capture from all nodes, hubs were installed in the topology. Port mirroring could also have been used.

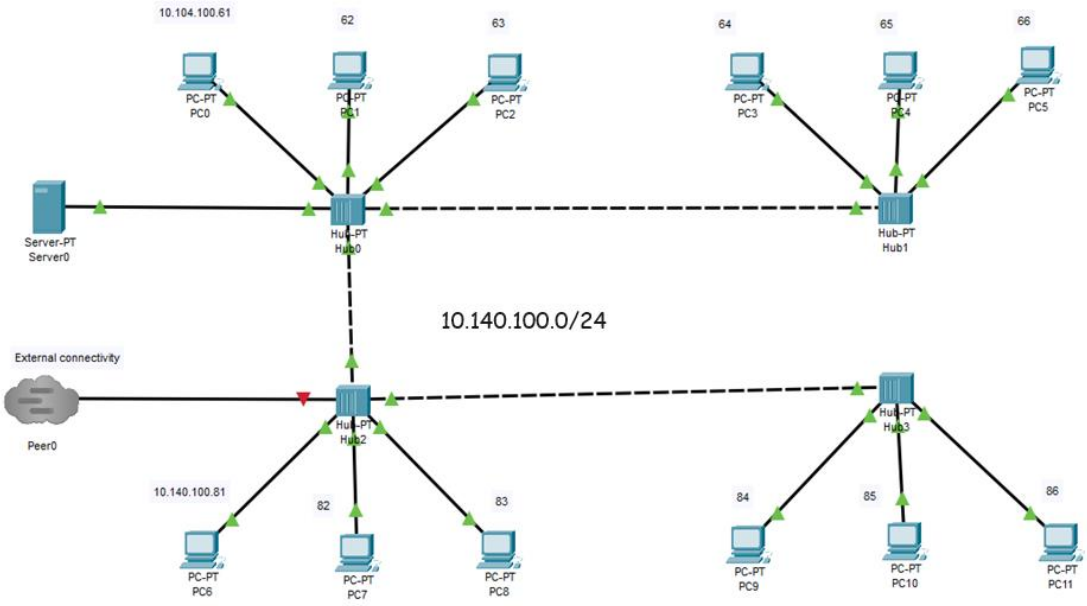

Figure 2 – Topology for capture1-10-11-23.pcapng

Publications associated with these datasets or the testbed

1) Hartpence, Bruce, and Andres Kwasinski. "A Convolutional Neural Network Approach to Improving Network Visibility." 2020 29th Wireless and Optical Communications Conference (WOCC). IEEE, 2020.

2) Hartpence, Bruce, and Andres Kwasinski. "Considering the Blackbox: An Investigation of Optimization Techniques with Completely Balanced Datasets of Packet Traffic." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

3) Hartpence, Bruce, and Andres Kwasinski. "Combating TCP Port Scan Attacks Using Sequential Neural Networks." 2020 International Conference on Computing, Networking and Communications (ICNC). IEEE, 2020.

4) Hartpence, Bruce, Kwasinski, A., “Fast Internet Packet and Flow Classification Based on Artificial Neural Networks”, IEEE Southeastcon, 2019.

5) Hartpence, Bruce, “Performance Evaluation of Networks with Physical and Virtual Links”, IEEE Global Information Infrastructure and Networking Symposium (GIIS), 2015.