EMIST Network Traffic Digesting (NTD) Tool Manual (Version I)

J. Wang, D.J. Miller and G. Kesidis

CSE & EE Depts, Penn State

Table of Contents

1. Overview

The EMIST NTD (Network Traffic Digesting) Tool is an off-line network traffic analysis tool capable of analyzing both TCPDUMP [1] and Cisco NetFlow [2] export format traces in Windows. The EMIST NTD tool can detect the significant clusters, i.e., clusters whose traffic is greater than a threshold (either in terms of packet number or bytes) that is user-specified. The thresholds can be specified for in a unidimensional fashion (for source IP, destination IP, source port, destination port or protocol) and also in multidimensional fashion for the five-tuple. The EMIST NTD tool is functionally similar to the AutoFocus tool in [3]; however, it is significantly more computationally efficient. The detailed algorithms for both unidimensional and multidimensional clustering processes are described in [4].

2. Getting Started

2.1. System Requirements

The EMIST NTD Tool only runs on Windows XP platform, and it is recommended to run the tool with 1.0 GHz or higher CPU and 256MB or higher memory.
The EMIST NTD Tool may need 15 seconds for Cisco NetFlow version 5 format input, and 20 seconds for simplified TCPDUMP format input, both tested for one-hour AucklandIV trace files [5] with Intel 3.0 GHz CPU (50% usage) and 1GB memory. The output digesting report will be saved automatically in the same folder as your input trace file. So please be patient when running the tool.

2.2. Installation

Simply download the file 'NTD.exe' from http://emist.ist.psu.edu/download_mining.html and run it on Windows XP platform.

3. Input Parameters

The EMIST NTD Tool version 1 only supports the simplified TCPDUMP format (Appendix I) and Cisco NetFlow version 5 format (Appendix II) inputs. Users need to input four parameters: the format of input trace file ('T' means TCPDUMP, 'N' means NetFlow version 5); the whole path of the input trace file; the report type ('B' means byte report, 'P' means packet report); the 'significant' threshold (a decimal fraction between 0 and 1). Please refer to the given example for details.

4. Output Report

The EMIST NTD Tool output report file will be saved automatically in the same folder as the input trace file with the name 'Report.txt'. The digesting report includes three parts: traffic characteristic introduction, unidimensional digesting report, and multidimensional digesting report.

4.1. Traffic Characteristic Introduction

This part includes the information of the starting time (EST time) of the input trace, the time interval of the input traffic, total bytes and number of packets in the traffic, and the threshold (user-specified) for digesting.

4.2. Unidimensional Digesting Report

This part includes five separate reports, one each for the following dimensions: source IP, destination IP, source port, destination port, and protocol. In order to present a concise, manageable report, only the compressed significant clusters [4] of each dimension are listed in decreasing order of their byte/packet percentage (of the total bytes/packets). The columns 'From' and 'To' list the starting time and the ending time, both relative to the starting EST time in part 1, for each cluster.

Notation Description:

*: This dimension is not used in defining the flow, i.e., the flow includes all of the possible values in this dimension;

Slash (/) in IP dimensions: network mask of IP addresses

'low port', 'high port' in port dimensions: 'low port' means the port group of port number less than 1024; 'high port' means the port group of port number larger than 1023

4.3. Multidimensional Digesting Report

Similar to the unidimensional report, only the compressed significant clusters are listed here in decreasing order according to their percentage value.

5. References

[1] TCPDUMP and LIBPCAP. *http://www.tcpdump.org/.*

[2] Cisco NetFlow Export Datagram Format. *http://www.cisco.com/univercd/cc/td/doc/product/rtrmgmt/nfc/nfc_3_0/nfc_ug/nfcform.htm#11770.*

[3] "Automatically inferring patterns of resource consumption in network traffic", C. Estan, S. Savage, and G. Varghese, SIGCOMM2003.

[4] "Efficient Mining of the Multidimensional Traffic Cluster Hierarchy for Digesting, Visualization, and Modeling", J. Wang, D.J. Miller, and G. Kesidis, pending.

[5] *Waikato Applied Network Dynamics Research Group*. Auckland University data traces. *http://wand.cs.waikato.ac.nz/wand/wits/.*

Appendix I : Simplified TCPDUMP datagram format

The simplified TCPDUMP datagram format only contains the useful information, which is included in most TCPDUMP traces, for the EMIST NTD Tool. So if your TCPDUMP trace file is in another format or arranged in another order, please extract the useful information from the original file according to the following format definition.

The simplified TCPDUMP datagram format has no header, so please remove the header from the original trace file. In the simplified TCPDUMP datagram format, there are only 24 bytes for each datagram record, they are: Nanosecond (4 bytes), Second (4 bytes), Source IP Address (4 bytes), Destination IP Address (4 bytes), Source Port Number (2 bytes), Destination Port Number (2 bytes), Datagram Length (2 bytes), Upper-layer Protocol (1 byte) and TCP Flag (it is always zero for UDP datagram) (1 byte).

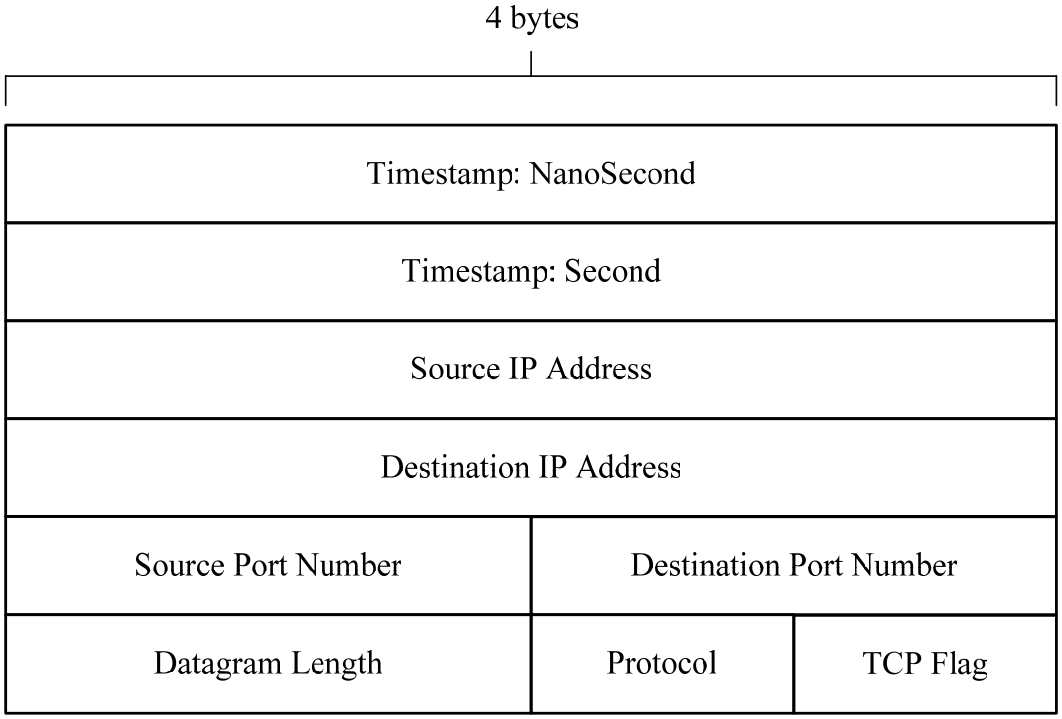

Figure 1. Simplified TCPDUMP datagram format

Appendix II: Cisco NetFlow Export Datagram Format (Version 5)

The EMIST NTD Tool version I also supports Cisco NetFlow version 5 datagram format. The header format and flow record format are shown in Table 1 and Table 2 [2].

Table 1. Cisco NetFlow Version 5 Header Format

Bytes	Contents	Description
0 – 1	Version	NetFlow export format version number
2 – 3	Count	Number of flows exported in this packet (1-24)
4 – 7	SysUptime	Current time in milliseconds since the export device booted
8 – 11	Unix_secs	Current count of seconds since 0000 UTC 1970
12 – 15	Unix_nsecs	Residual nanoseconds since 0000 UTC 1970
16 – 19	Flow_sequence	Sequence counter of total flows seen
20	Engine_type	Type of flow-switching engine
21	Engine_id	Slot number of the flow-switching engine
22 – 23	Reserved	Unused (zero) bytes

Table 2: Cisco NetFlow Version 5 Flow Record Format

Bytes	Contents	Description
0 – 3	srcaddr	Source IP address
4 – 7	dstaddr	Destination IP address
8 – 11	nexthop	IP address of next hop router
12 – 13	input	SNMP index of input interface
14 – 15	output	SNMP index of output interface
16 – 19	dPkts	Packets in the flow
20 – 23	dOctets	Total number of Layer 3 bytes in the packets of the flow
24 – 27	First	SysUptime at start of flow
28 – 31	Last	SysUptime at the time the last packet of the flow was received
32 – 33	srcport	TCP/UDP source port number or equivalent
34 – 35	dstport	TCP/UDP destination port number or equivalent
36	pad1	Unused (zero) bytes
37	tcp_flags	Cumulative OR of TCP flags
38	prot	IP protocol type (for example, TCP = 6; UDP = 17)
39	tos	IP type of service (ToS)
40 – 41	src_as	Autonomous system number of the source, either origin or peer
42 – 43	dst_as	Autonomous system number of the destination, either origin or peer
44	src_mask	Source address prefix mask bits
45	dst_mask	Destination address prefix mask bits
46 – 47	pad2	Unused (zero) bytes