



Collecting and Formatting Data For the ExaSphere Analysis Program

University of South Carolina Advanced Solutions Group

1 Introduction

This document will explain how to collect data and format it for input into the Exasphere Network Analysis Engine. The method explained in this document should not be considered the only way obtain properly formatted data; it should be thought of as a suggestion that may modified in order to work optimally in the given situation. Once the data file has been created and properly formatted, the Network Analysis Engine operates the same way on all platforms.

2 What Information Is Needed?

Currently the Network Analysis Engine is set up to accept the following pieces of data about each packet:

- Timestamp (formatted as the number of milliseconds since the Unix Epoch began: January 1, 1970)
- Source IP Address
- Destination IP Address
- Destination Port
- Size of the packet (in bytes)

3 Capturing The Data

It should be noted that on modern switched networks, only broadcast packets and packets intended to go to your computer actually pass to your network adapter. So in order to be able to capture traffic intended for other machines on the network, you need access to the switch. If you are not the network administrator on your network, you will need to seek the assistance of that person for your network.

There are many programs that can capture all network traffic that arrives at the network adapter on your computer. Two that are readily available for multiple platforms and free to the public are

- tcpdump: www.tcpdump.org (a Windows version is at www.winpcap.org/windump/)
- ethereal: www.ethereal.com

For the sake of this discussion, we will use tcpdump. The following command generates a dump file that includes the data that is needed for the Network Analysis Engine. This command will run until you interrupt it and then write the dump to the file dump.txt.

```
tcpdump ip -nn -tttt -e > dump.txt
```

The command line switches used above are

- `ip`: Capture only IP traffic
- `-nn`: Do not resolve the IP addresses or port service names
- `-tttt`: Include the time and date
- `-e`: Include link-level header information

and result in an output file that contains lines like the following:

```
05/10/2006 20:42:46.175485 0:c8:9f:1f:75:b7 0:14:11:31:d4:c2 0800
134: 192.168.0.46.22 > 192.168.0.99.60300: P 1234:1314(80)
ack 1029 win 7120 (DF)

05/10/2006 20:42:46.175652 0:90:17:74:81:fe 0:c0:7f:1f:75:b7 0800
172: 192.168.0.52.53 > 192.168.0.46.33205: 32947 NXDomain*
0/1/0 (130)

05/10/2006 20:42:46.176399 0:11:11:61:d4:c2 0:b0:9f:1f:75:b7 0800
422: 192.168.0.99.60300 > 192.168.0.46.22: P 1029:1397(368)
ack 1314 win 64075 (DF) [tos 0x5c]
```

It is obvious that this dump contains much more information than is needed by the analysis engine. The next step is to pull out just the information that needed and format it properly for input into the engine.

4 Formatting The Data

Each line of the Network Analysis Engine input data file needs to be in the following format:

```
timestamp    source ip    destination ip    destination
port    packet size
```

Each line of the input file represents a single packet. The fields are separated by any amount of whitespace and there should be no whitespace within a field. In general, you will need to process the dump file to create an input file of this format. There are many ways to accomplish this task, but writing a script in a good text parsing language is probably the easiest thing to do. The file **fix_dump.pl** included with this distribution is a perl implementation of that procedure. Once the dump file has been processed to create the input file, you are ready to use the Network Analysis Engine.

5 Using The Network Analysis Engine

In order to use the Network Analysis Engine, you must have a Java Runtime Environment (JRE) installed. The JRE is a free piece of software available from <http://java.sun.com>.

5.1 Create The Input Files

Make a file with a list of all of the IP addresses on the network you wish to analyze and store them in a text file. This is your valid ips file. Make sure the input data file is properly formatted and then start up the Network Analysis Engine.

5.2 How To Run The Engine:

The first thing that you need to do is to set the parameter variables by choosing "Set Variables" from the Options menu. You may wish to experiment with these values to give you the best results. The parameters are

- Time interval: The duration of one analysis. The engine stops reading new data at the end of each interval and calculates the entropies. It repeats this process until all of the data has been read. The default time interval is 1 minute.
- Lambda: A value that determines the effect of connectivity. The value of lambda determines how sensitive the analysis is to more distant connections (in terms of intermediate nodes). The higher the value, the more sensitive. The lower the value, the less sensitive. Lambda needs to be big enough to properly include distant nodes, but a value of lambda that is too large will result in an error. If an error occurs, the engine will produce a beep and the analysis will stop. The default value for lambda is 1.
- Power Term: A value that determines the connection depth of nodes that the Network Analysis Engine considers. For example, if A-B ("-" means "is connected to"), B-C, C-D, and D-E, a power term of 3 means that the engine considers A to be connected to B, C, and D, but not E. A large power term leads to a more accurate analysis of the network but increases processing times. The default power term is 3.

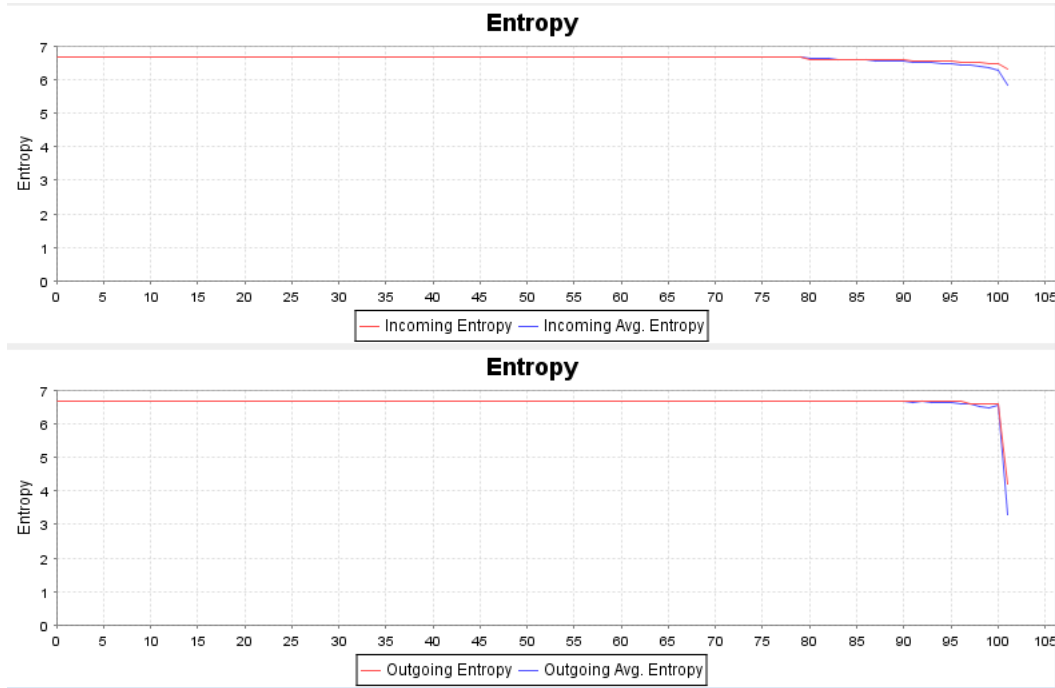


Figure 1: Entropy Profile

Once all of the variables are set, you can calculate the entropy by choosing "Calculate Entropy" from the File menu. This opens a dialog box which allows you to choose three files. The first is the "list file" which is the file with the list of valid ip addresses on your network. The second is the data file which is the formatted network dump. The third file is for saving the average behavior of the network entropies for the selected input files. When the calculation is done, the program saves the average entropy behavior and the deviation to this file. Click the Entropy panel once to make the plots appear. (see Figure 1).

Now you can see how the network entropy graph changes in time by choosing "Calculate Entropy with R-Squared" from the File menu. This opens a dialog box which asks for three files as in the previous step. The only difference is that you should select for the third file the average entropy file that you saved in the previous step. The program then displays the fluctuation (in the "Fluctuation" panel) of network entropy based on the average behavior. Click on the Entropy panel once to make the plots appear

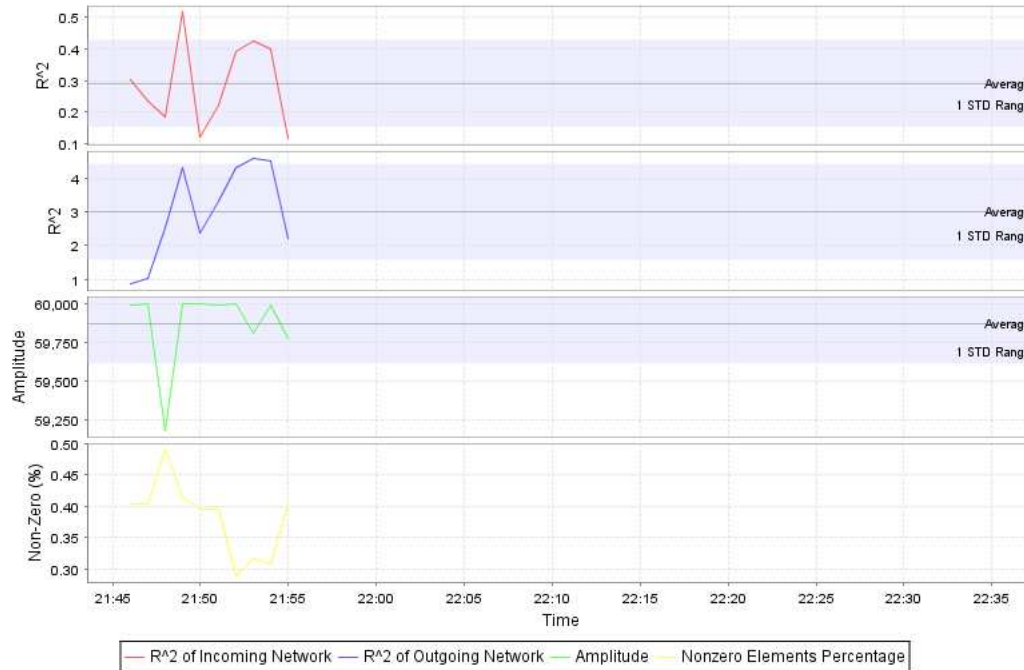


Figure 2: R-Squared Plot

(See Figure 2).

You can stop the calculation by choosing "Stop Calculation" from File menu. When you do this, the Network Analysis Engine will finish the current calculation and then stop so the engine may not stop immediately.

5.3 How To Interpret The Plots

For the Entropy Calculation, you will see entropy graphs of each time interval as the analysis for that interval completes. This process will continue until the engine reaches the end of the data set. The entropies of the incoming and outgoing traffic are plotted with the outgoing entropy sorted in descending order and the incoming entropy sorted by node in the same order as the sorted outgoing entropy. A low entropy value for a nodes indicates a high level of activity.

For the Entropy Calculation With R-Square, you can see the entropy plots and the fluctuation plots. The entropy plots change as each interval analysis completes as before. The Fluctuation panel shows how the entropy

for each interval differs from the average behavior. The R-Squared values for the outgoing and incoming traffic are high for a large deviation from the average and low for a small deviation.

The Amplitude is a measure of traffic volume. A high value for the Amplitude means that there was a high volume of network activity during that time.

By clicking any point on the graph, you will see a line appear that shows the R-Squared and Amplitude values at that time. By right-clicking any point on the graph, you can see the outgoing and incoming entropy plots for that time.

6 Limitations

The Network Analysis Engine currently has the following limitations:

- The maximum number of unique IP addresses in the valid ips file is 1000.
- The packet size in the input file must be an integer.