Homework #2

OBJECTIVE

This homework assignment is designed to give you some hands on expertise with a more sophisticated network measurement tool. You will use the Argus flow analysis tools to study information available in flow traces.

GUIDELINES

This is not a group assignment. You should work on this homework individually and write a report on your own. It is not acceptable to just look over your classmate's shoulder and gather the answers for your report. It is fine to discuss the homework with others, but you must do all the work involved in such a manner as to understand the exercises. In particular, if your answers are substantially identical to those of another student's answers, both of you will be considered for academic integrity violations.

The points associated with each problem are listed after the problem. Most will be graded in an "all-or-nothing" manner, so show all your work for full credit.

Write a report and submit it to Gradescope. Late submissions will not be accepted. No exceptions. The report should be in PDF format.

Get started early. The commands you will be using take considerable time to execute. If you start the night before the deadline, you may not finish.

The TAs tell me that the best way to complete this homework (from their experience when they were in the class) is to spend some time reading the Argus documentation thoroughly -- and then attempt the questions. If you read a question and try to go into the documentation searching for the syntax for that particular question, you can get pretty confused because you haven't seen the (sometimes subtle) distinctions between the various commands and command-line arguments.

THE FLOW DATA

As you learned in the Network Measurement lecture, a flow is a unidirectional stream of packets between the same source and destination. Generally, 7 fields must be identical among packets in the flow: Source IP address, Destination IP address, Source Port number, Destination Port number, Protocol value (i.e. TCP, UDP, other), Type of Service byte and Input logical interface. Additional data about the flow is included in the flow record; usually the byte and packet counts for the flow. We will see that some flow tools also perform TCP header inspection and can therefore provide additional information per flow.

You will have access to flow data from a network defense exercise. The data itself can be found at **`/afs/andrew/course/14/740/hw2/cdx.arg`**. You may wish to make a symbolic link to the data file to make it easier to access. Don't simply copy the file, as it is rather massive.The following command will make such a symbolic link in the current directory (i.e. from your directory):

```
ln -s /afs/andrew/course/14/740/hw2/cdx.arg cdx.arg
```

All of the needed tools are available in **`/afs/andrew/course/14/740/bin/bin`**[1] and should be accessed remotely from **`unix.andrew.cmu.edu`**. You should probably add this directory to your PATH with:

```
export PATH=/afs/andrew/course/14/740/bin/bin:$PATH
```
[2]

If you wish to build the argus tools for your own machine, you are welcome to do so. You can find the tools, documents and more information at http://qosient.com/argus/.

[1] Yep, bin appears twice. Don't ask why.

[2] If you aren't using bash as your shell, the syntax is slightly different.

THE TOOLS

For this lab, we'll be using the argus flow tool to examine the flow data. Argus was originally developed at CERT in 1993 and consists of two tool sets:

- Argus server: used to capture and convert flow data
- Argus client: used to examine flow data

The main benefit of Argus is that it constructs flows in a bidirectional manner. An example of an Argus flow can be seen below:

```
  StartTime        Flgs   Proto SrcAddr    Sport  Dir   DstAddr  Dport

   07:44:52.549815  e      6    10.1.20.5.56807   ->   10.1.60.5.53
```

The Argus client suite consists of the following tools:

- **`ra:`** reads argus data, filters the records based on a filter expression, and prints the contents of the records.
- **`racount:`** prints out various counts from the data in an argus file.
- **`racluster:`** clusters the records based on a specified flow key criteria.
- **`rabins:`** aggregates data to a set of bins, or slots. Mainly used to aggregate data on a time series.
- **`rasort:`** sorts the records based on the specified criteria.

The argus client tool set contains many more executables that are not relevant to this homework. Note: You may find the following unix commands to be very helpful: **`wc`**, **`head`**, **`sort`**, **`col`**. Also, the ability to redirect the terminal output into a file (perhaps for later consumption by a spreadsheet for graphing?) is very helpful. Redirection of a command is accomplished by using the redirection operator (**>**) followed by a filename. For instance, **`command > filename`** will place all of the text that would have been printed to the terminal window when **`command`** was executed and put it into a file named **`filename`**.

FOLLOW THESE STEPS:

1) Login to **`unix.andrew.cmu.edu`** using ssh.

2) Read some of the argus tool documentation (i.e. man pages) at http://qosient.com/argus/manuals.shtml. You might also find the examples in the Argus wiki to be helpful (http://nsmwiki.org/index.php?title=Argus#Examples). Experiment with the 5 argus tools mentioned on the previous page.

3) Devise sequences of argus commands to answer the questions that follow. For each question, make sure to document the command line you used to generate the question. You may have to discover other information (like specific port numbers). Make sure to also document how you discovered that information.

You will be graded on the following criteria:

- Conclusions are insightful and fully supported
- All assumptions are reasonable and clearly stated
- Procedure used to solve the problem is fully explained
- Results are complete and accurate
- Explanations are clear, concise and use proper English (spelling, grammar and usage)

The more understanding and research of the network you show, the better grade you will get.

Answer the following questions about the cdx.arg dataset:

1. (5 points) How many flows are TCP? UDP? Others?

2. (5 points) What percentage of the total traffic (by bytes) was DNS? What about by packet?

3. (15 points) What are the traffic loads (by bytes) for each 1 hour throughout the capture? Provide a graph, not a list.
   Hint: You can redirect the output of the appropriate argus command to a file (command > output.csv) and open it with your favorite spreadsheet software.

4. (10 points) What IP address is receiving the most traffic? How much data is it receiving? How much is it sending? You will need to define "most traffic" for yourself -- and make sure to discuss and justify your decision.

5. (10 points) What are the top 2 services (by bytes) running on this machine (the one you discovered in question #4) ? How much traffic is each of them generating?

6. (5 points) What IP address is sending the most traffic to the IP from question #4 on the first port from question #5?

7. (15 points) Read up on "TCP SYN Flood attack" in your textbook and Wikipedia. Looking at the traffic between these 2 IP addresses on the port from question #5, does this traffic resemble a SYN flooding attack? Explain why or why not, and justify your answer using the corresponding argus tools.

8. (10 points) How many active mail servers (SMTP) can you identify on this network and what are their IP addresses?

9. (25 points) Find something interesting in the data and examine it (i.e. make up your own question). Be creative and search out something that flow tools might typically be used to check. You will be graded on how creative / interesting your question is and how thoroughly you answer it. In particular, ensure your question is not a simple modification of one of the questions I asked above (i.e. "How much traffic is FTP?" is a poor question to examine).