Strip Plots: A Simple Automated Time-Series Visualization

Sid Faber
sfaber@cert.org
Overview

Motivation & Goals
Sample Output
The Basics
Special Features
NO WARRANTY

THIS MATERIAL OF CARNEGIE MELLON UNIVERSITY AND ITS SOFTWARE ENGINEERING INSTITUTE IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

Use of any trademarks in this presentation is not intended in any way to infringe on the rights of the trademark holder.

This Presentation may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

This work was created in the performance of Federal Government Contract Number FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. The Government of the United States has a royalty-free government-purpose license to use, duplicate, or disclose the work, in whole or in part and in any manner, and to have or permit others to do so, for government purposes pursuant to the copyright license under the clause at 252.227-7013.
Motivation and Goals
Caveat

This is analyst code, not engineering code

• Your mileage may vary
Motivation

Support network profiling for Situational Awareness

• I know most of what’s on my network
  • Based exclusively on past observations
• I can filter / categorize out routine traffic
• What can I do with “leftover” traffic?
  • Is this something new to add to my profile?
  • Is something odd happening on my network?
Example
Goals

Nearly self-maintaining network profile

- Batch processing
- Email delivery
- Quick triage of “leftovers”
  - Add to my profile?
  - Something odd?
  - <5 minutes per report
- Self-sufficient description: *No Additional Explanation Necessary*
<table>
<thead>
<tr>
<th>sIP</th>
<th>dIP</th>
<th>dPort</th>
<th>dPort</th>
<th>pktS</th>
<th>bytes</th>
<th>flags</th>
<th>sTime</th>
<th>dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33.6.97</td>
<td>181.113.55.11</td>
<td>63220</td>
<td>80</td>
<td>6</td>
<td>104</td>
<td>F A</td>
<td>2009/11/17T00:00:01.236</td>
<td>0.071</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>0.33.6.97</td>
<td>63220</td>
<td>80</td>
<td>6</td>
<td>104</td>
<td>F A</td>
<td>2009/11/17T00:00:01.236</td>
<td>0.000</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>94.145.72.124</td>
<td>50151</td>
<td>6</td>
<td>8</td>
<td>9016</td>
<td>S PA</td>
<td>2009/11/17T00:00:07.233</td>
<td>0.232</td>
</tr>
<tr>
<td>92.4.182.32</td>
<td>181.113.55.11</td>
<td>1067</td>
<td>80</td>
<td>6</td>
<td>2012</td>
<td>PA</td>
<td>2009/11/17T00:00:09.827</td>
<td>2.125</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>92.4.182.32</td>
<td>1067</td>
<td>80</td>
<td>6</td>
<td>2532910</td>
<td>PA</td>
<td>2009/11/17T00:00:09.840</td>
<td>1.713</td>
</tr>
<tr>
<td>92.4.182.32</td>
<td>181.113.55.11</td>
<td>1056</td>
<td>80</td>
<td>6</td>
<td>201868</td>
<td>PA</td>
<td>2009/11/17T00:00:09.841</td>
<td>2.684</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>92.4.182.32</td>
<td>1056</td>
<td>80</td>
<td>6</td>
<td>2430819</td>
<td>PA</td>
<td>2009/11/17T00:00:09.842</td>
<td>2.355</td>
</tr>
<tr>
<td>139.173.53.191</td>
<td>181.113.55.11</td>
<td>9429</td>
<td>80</td>
<td>6</td>
<td>140</td>
<td>R A</td>
<td>2009/11/17T00:00:12.299</td>
<td>0.000</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>184.199.238.201</td>
<td>33559</td>
<td>80</td>
<td>6</td>
<td>3341996</td>
<td>FS PA</td>
<td>2009/11/17T00:00:12.866</td>
<td>1.222</td>
</tr>
<tr>
<td>184.199.238.201</td>
<td>181.113.55.11</td>
<td>33559</td>
<td>80</td>
<td>6</td>
<td>281878</td>
<td>FS PA</td>
<td>2009/11/17T00:00:12.866</td>
<td>1.222</td>
</tr>
<tr>
<td>181.113.55.11</td>
<td>250.227.45.82</td>
<td>54162</td>
<td>80</td>
<td>6</td>
<td>341713</td>
<td>FS PA</td>
<td>2009/11/17T00:00:14.773</td>
<td>30.779</td>
</tr>
<tr>
<td>250.227.45.82</td>
<td>181.113.55.11</td>
<td>54162</td>
<td>80</td>
<td>6</td>
<td>75752</td>
<td>FS PA</td>
<td>2009/11/17T00:00:14.773</td>
<td>30.779</td>
</tr>
</tbody>
</table>

![Graph with data points](image-url)
Sample Output
One Hour of Standard HTTP Traffic
More http: can you spot the beacon?
Conficker
30-byte UDP packets
More web...

23 pages: Review traffic for 230 sip/dip pairs in <1 minute
The Basics
Basic script workflow

Find the top talkers
Count in & out traffic for each talker
Plot traffic for each
Compile plots into a single .pdf
Dependencies

SiLK toolset
gnuplot
ghostscript (gs)
Python (but not pySiLK)
Usage: stripplot.py [options] FILE

Creates a strip-plot of the most significant traffic contributors within a raw SiLK data file

FILE   (required) The binary SiLK file to analyze

Options:
--binsize  Default counting bin size (seconds) (def=auto)
--bottomleft Bottom left tag; allows %(substitution)s; use '-' for none (def=-)
--bottommiddle Words on the bottom middle of the page, allows %(substitution)s; use '-' for none (def=FOUO)
--bottomright Words on the bottom right of the page, allows %(substitution)s; use '-' for none (def=Page (%(pagecount)s) of Page (%(page)s))
--count Number of plots to output (def=5)
--endtime  Plot end time, YYYY/MM/DDTHH:MM:SS (def=auto)
--fields  rwuniq-style list of fields to group IN traffic on, or '*' for automatic (NOT ALL FIELDS WORK) (def=sip,dip)
--flags Include this option to add a plot of TCP flags to the strips (def=0)
--help  Print this output (def=)
--types Inbound and outbound types; these are used to make sure the IN address is on the top plot; must be in the form [in-type/out-type]; unspecified types work fine but either address may end up on the top. (def=in/out,inweb/outweb,inicmp/outicmp)
--pdffilepath PDF final output file (def=tmp.pdf)
--plotfile  Temporary gnuplot script file to create (def=tmp.plot)
--plotsperpage Number of plots per page (def=10)
--prefilter rwfilter expression to apply to flow file before selecting what to plot. NOTE: this filter is NOT applied to the trends themselves, only to the selection routine (def=--proto=0-)
--psfilepath Post-script file to generate (def=tmp.ps)
--selectionval Choose the top [count] combinations to plot based on this value; must be either 'bytes', 'packets', 'flows' or 'none'); if 'none' then output is in rwuniq (random) order (def=bytes)
--topleft Words on the upper left of the page, allows %(substitution)s; use '-' for none (def=-)
--topright Words on the upper right of the page, allows %(substitution)s; use '-' for none (def=-)
--trendline Highlighted and dotted trendline to add to plot; f for flows, p for packets (def=b)
--starttime Plot start time in YYYY/MM/DDTHH:MM:SS format (def=auto)
--maintitle Title for this plot, allows %(substitution)s; use '-' for none (def=auto)
--verbose Print out debugging info, use twice for more info and to print debugging info on the plot itself (def=0)

Fields with string substitution support the following:
%(page)i  Current page number
%(date)s  Date the report was printed
%(time)s  Time the report was printed
%(pagecount)i  Total number of pages in the report
%(setting)s Any of the report configuration settings (run with -v option to see settings)
An individual strip

**Internal** IP address and/or port, or a "***" for all addresses

Red shaded areas show traffic byte volume generated internally and sent to the external address and/or port

**External** IP address and/or port, or a "***" for all addresses

Blue shaded areas show traffic byte volume entering the network from the external address and/or port

Byte volume magnitude label (Kbytes per second); same for both top and bottom plots

Timeline labels

Red shaded areas show traffic byte volume generated internally and sent to the external address and/or port

Blue shaded areas show traffic byte volume entering the network from the external address and/or port

Timeline labels

**Internal** IP address and/or port, or a "***" for all addresses

Red shaded areas show traffic byte volume generated internally and sent to the external address and/or port

Byte volume magnitude label (Kbytes per second); same for both top and bottom plots

Timeline labels

**External** IP address and/or port, or a "***" for all addresses

Blue shaded areas show traffic byte volume entering the network from the external address and/or port

Timeline labels
The Full Page

- Time scale/formatting courtesy of gnuplot
- Consistent timeline throughout
- Each strip has an independent magnitude, but top and bottom are the same magnitude

Labels with substitutable parameters
--fields=[rwuniq field set] defines groupings

--fields=sip/dip (default): group on client/server pairs

--fields=dip: group on dest (my) address

--fields=sip,dip,dport: group on external services

--fields=sip,dip,sport,dport: group on TCP sessions
Q: How does it choose what strips to plot?

A: It uses `rwuniq` to group records, then chooses the largest by byte volume:

```
rwuniq [file] \n   --fields=[field list] \n   | sort -nr -k [bytes column]
```
Tweaking the selection criteria

Sometimes you want to show the top packets or flows

- Repeated failed connection attempts
- Bot phone-home

```
--selectionval=flows|packets
```

- Sort by top flow count or packet count
selectionval example

Select by Bytes

Select by Flows
Other Common Options

--count
  - Defines how many strips to plot, defaults to 5

--prefilter
  - Filter out the .rw file before plotting it
Trend lines

--trendline=f [or p]

- Adds a highlighted dotted line for flows (or packets)
- No labeling, no magnitudes, not symmetric
- Good for drawing out low-volume data

The trend line shows ACK packet counts generally match data packets
Bin size

Points on the time axis:

One hour defaults to 1 second bins (3600 points):

5 second bins makes little difference:

At 10 seconds (360 bins), the picture gets fuzzy:
Starttime, endtime

Zoom in on a particular time frame:

```
./stripplot.py --fields=dip sample.anon.rw
--starttime=2009/11/17T00:00:00 --endtime=2009/11/17T00:10:00
--trendline=p sample.anon.rw
```
Special Features
Automatically selecting fields

--fields=*  

- Selects the best combination for 2, 3, 4 or 5 of sip, dip, sport, dport, proto
- For example, looking at all traffic for one network:
  *:80  ->  230.28.41.254:*  
  *:1935 ->  230.28.41.254:*  
  *:*   ->  230.28.136.207:80  
  *:4500 ->  *:4500  
  *:*   ->  230.28.229.52:25  
  45.178.111.132:61296  ->  230.28.41.74:56066  
  239.213.117.254:8080  ->  230.28.41.254:53474  
  *:*   ->  230.28.229.75:25  
  45.178.111.132:50769  ->  230.28.41.74:55722  
  *:443  ->  230.28.41.254:*
Auto-select Example

- Proxied / NAT'd Web Traffic
- Proxied / NAT'd Shockwave Traffic
- My largest web server
- IPSec NAT Traversal
- Incoming eMail
Auto-select: how does that work?

Simple, but slow. Find the top value for rwuniq on:

- sip,sport,dip,dport
- sip,sport,dip
- sip,sport,dport
- sip,sport
- sip,dip,dport
- sip,dip
- sip,dport
- sport,dip,dport
- sport,dip
- sport,dport
- dip,dport
Plotting TCP Flags

--flags

- Display points for TCP flags
- Works fine, just can’t find a good generic use case

Activity: Grey bar

```
./stripplot.py --fields=sip,dip 
    --flags sample.anon.rw
```
Plotting TCP Flags

--flags

- Display points for TCP flags
- Works fine, just can’t find a good generic use case

```
./stripplot.py --fields=sip,dip \
--flags sample.anon.rw
```

SYN Packets (outside, black)
Plotting TCP Flags

--flags

- Display points for TCP flags
- Works fine, just can’t find a good generic use case

```bash
./stripplot.py --fields=sip,dip \
--flags sample.anon.rw
```

FIN Packets (middle, black)
Plotting TCP Flags

--flags

- Display points for TCP flags
- Works fine, just can’t find a good generic use case

./stripplot.py --fields=sip,dip
--flags sample.anon.rw

RST Packets (middle, blue or red)
Verbose output, -v

$ ./stripplot.py --fields=sip,dip --flags -v sample.anon.rw
Found 17.117.24.230:* -> 181.113.55.11:*
Found 84.101.142.241:* -> 181.113.55.11:*
Found 5.73.229.27:* -> 181.113.55.11:*
Found 159.133.127.154:* -> 181.113.55.11:*
Found 195.208.192.99:* -> 181.113.55.11:*

# Settings:
# binsize  1
# bottomMargin 0.08
...
# usableWidth 0.9
# verbose  1
# workfile  sample.anon.rw
(000): Page 001, Plot 000
(001): Page 001, Plot 001
(002): Page 001, Plot 002
(003): Page 001, Plot 003
(004): Page 001, Plot 004
Very Verbose Output, -vv

Echos all rw* commands

Adds lots of info to the .pdf output
Overriding default types

--types=in/out,inweb/outweb

- It’s OK if the type doesn’t actually exist (i.e., multiple installations)
- Has to match rwcut type field
Open Issues

Better error checking
  • Most inputs are passed directly to the rw* tools
  • Occasional errors trying to plot empty data sets

Doesn’t work well for transit traffic
  • Assumes “in” and “out” traffic

Problems if you have outbound traffic only
  • Trend selection doesn’t seem to work well

TCP flags
  • Tech is there, but visualization needs lots of improvement
Strip Plots: A Simple Automated Time-Series Visualization

Sid Faber
sfaber@cert.org