



Visual Analytics for Streaming Internet Traffic

Edward J. Wegman

George Mason University

Karen Kafadar

University of Colorado, Denver

Visual Analytics for Streaming Internet Traffic

◆ The following discussion is based on the following three papers

- Wegman, E. and Marchette, D. (2003) "On some techniques for streaming data: A case study of Internet packet headers," *Journal of Computational and Graphical Statistics*, 12(4), 893-914
- Marchette, D. and Wegman, E. (2004) "Statistical analysis of network data for cybersecurity," *Chance*, 17(1), 8-18
- Kafadar, K. and Wegman, E. (2004) "Visualizing 'typical' and 'exotic' Internet traffic data," Proceedings of COMSTAT2004.

Visual Analytics for Streaming Internet Traffic

◆ Introduction

◆ Visual Analytics

- Analysis versus Exploration

◆ Block Recursion and Evolutionary Graphics

- Waterfall plots
- Skyline plots
- EWMA plots

Four Stages of Data Graphics

1. Static Graphics

- ◆ Ed Tufte's books
- ◆ Trellis plots, scatterplot matrices, parallel coordinate plots, most density plots
- ◆ Most (paper) published materials
- ◆ Perhaps some color and anaglyph stereo

2. Interactive Graphics

- ◆ Data objects created, but underlying data untouched
- ◆ Think of data on server, graphics on client
- ◆ Brushing, saturation brushing, 3-D (stereoscopic) plots
- ◆ Rocking and rotation, Dan Carr's micromaps, cropping and cutting, linked views

Four Stages of Data Graphics

3. Dynamic Graphics

- ◆ Data that must be interacted with, not just client based
- ◆ Grand tour, multidimensional grand tour, recursive or dynamically smoothed density plots, dynamic smoothing including mode trees and mode forests, Dan Carr's conditioned chloropleth maps, pixel tours, cross corpora discovery

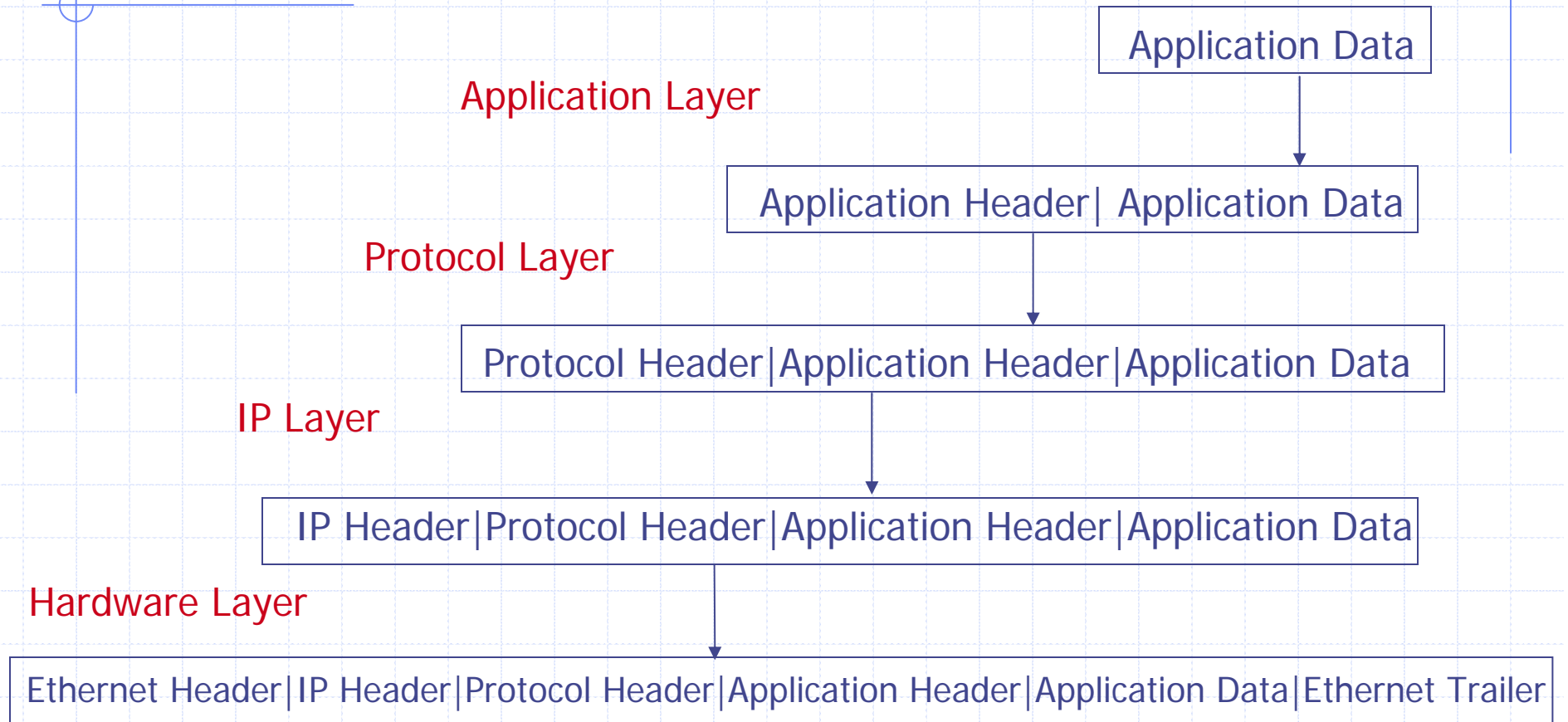
4. Evolutionary Graphics

- ◆ Fixed data sets that are evolving
 - ◆ Data Set Mapping
 - ◆ Iterative Denoising
- ◆ Streaming data
 - ◆ Recursion and Block Recursion
 - ◆ Visual Analytics
 - ◆ Waterfall, Transient Geographic Mapping, Skyline Plots

Visual Analytics for Streaming Internet Traffic - Types of Networks

- ◆ Class A – field1 identifies the network, fields2-4 identify the specific host
 - field1 is smaller than 127, e.g. 1.1.1.1
- ◆ Class B – field1.field2 identifies the network field3.field4 identifies the specific host, field3 sometimes used for subnet
 - Field1 is larger than 127, e.g. 130.103.40.210
- ◆ Class C- field1.field2.field3 identifies the network, field4 the host
 - E.g. 192.9.200.15

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing



Visual Analytics for Streaming Internet Traffic - Common Protocols

- ◆ TCP=Transmission Control Protocol
- ◆ UDP=User Datagram Protocol
- ◆ ICMP=Internet Control Message Protocol

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

| | | | | |
|------------------------|----------|-----------------|---------------------|-----------------|
| Version | Length | Type of Service | Total Packet Length | |
| Identification | | | Flags | Fragment Offset |
| Time to Live | Protocol | | Header Checksum | |
| Source IP Address | | | | |
| Destination IP Address | | | | |
| Options (if any) | | | | |

The IP Header

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

| | | | |
|-----------------------|----------|------------------|-------------|
| Source Port | | Destination Port | |
| Sequence Number | | | |
| Acknowledgment Number | | | |
| Length | Reserved | Flags | Window Size |
| Checksum | | Urgent Pointer | |
| Options (if any) | | | |

TCP Packet Header

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

◆ Some Flag Types

- ACK – used to acknowledge receipt of a packet
- PSH – data should be pushed to application ASAP
- RST – reset
- SYN – synchronize connection so each host knows order of packets
- FIN – finish the connection

Visual Analytics for Streaming Internet Traffic - TCP/IP Addressing

| HOST 1 | HOST 2 |
|--------------------------|----------------|
| SYN | SYN/ACK |
| ACK PSH PSH PSH | |
| | ACK PSH |
| ACK FIN | |
| | FIN/ACK PSH |
| ACK | |
| | FIN |
| FIN/ACK | |

Possible TCP Session

Visual Analytics for Streaming Internet Traffic - Ports

- ◆ There are some $2^{16} = 65,536$ ports for each host
 - Some standard services use standard ports
 - ◆ e.g. ftp – 21, ssh – 22, telnet – 23, smtp – 25, http – 80, pop3 – 110, nfs – 2049, even directv and aol have standard ports.
 - Unprotected (open) ports allow possible intrusion
 - ◆ Scanning for ports is a hacker attack strategy



Evolutionary Graphics from Streaming Internet Data

The major problem is to detect intrusions or unwanted events in streaming Internet traffic.

Evolutionary Graphics from Streaming Internet Data

1. Internet traffic is a prototypical example of streaming data and prefigures future streaming data types. I believe streaming data represents a fundamentally new data structure.
2. The papers mentioned above describe the basic protocols for Internet traffic. We look only at time stamps, destination IP, destination port, source IP, source port, number of bytes, number of packets, duration of session.
3. We ignore the data content of the packet, and seek to make inferences based only on the header data described above.

Evolutionary Graphics from Streaming Internet Data

1. Streaming data arrives as such a rate that it is impossible to store the data.
 - We collect 26 terabytes of Internet header data per year.
 - Naively, we look at a data item, update a recursive algorithm, and discard the data.
2. Some suggestions we have made include:
 - Recursive formulations of counts and moments
 - Pseudo-samples based on geometric quantization
 - Recursive formulations of kernel and adaptive mixture density estimators
 - Exponentially weighted moving averages including exponentially weighted kernel smoothers.

Evolutionary Graphics from Streaming Internet Data

But this talk is about **evolutionary graphics**

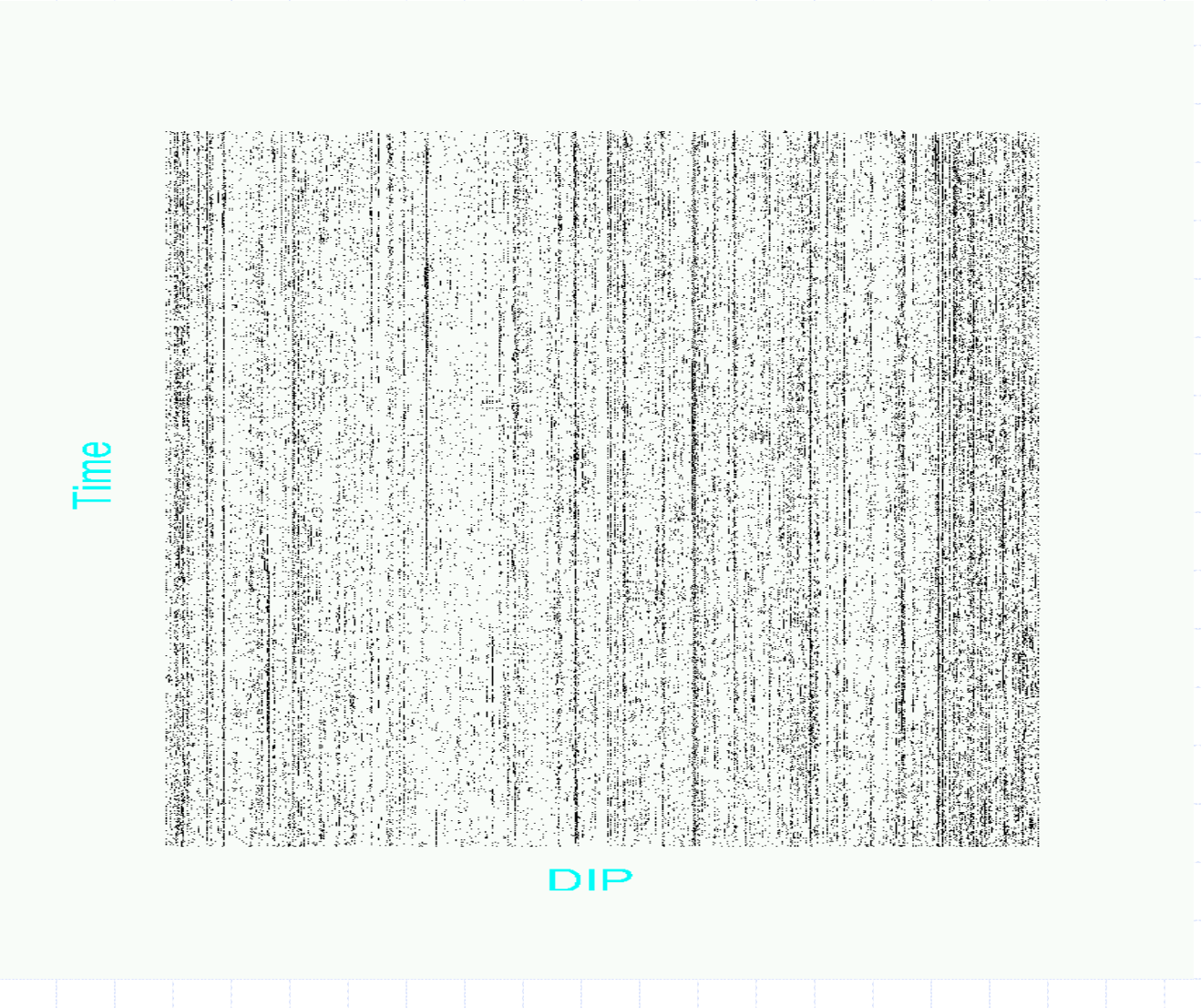
1. In the simplest framework, the idea is to accumulate data for a very small epoch (even instantaneously), plot the new data, and discard the old.
 - In practice for Internet traffic, the epoch may last for perhaps 10 milliseconds.
 - Our initial suggestion is a Waterfall diagram. The first epoch is plotted at the top of the graphic. As additional data are accumulated during the second epoch, the graphic for the first epoch is pushed down and the second epoch is now plotted on the top. This continues until perhaps 1000 epochs have passed. Thus the oldest epoch drops off the bottom of the page and the new replaces it at the top. The graphic evolves and is new every 10 seconds.



Evolutionary Graphics from Streaming Internet Data

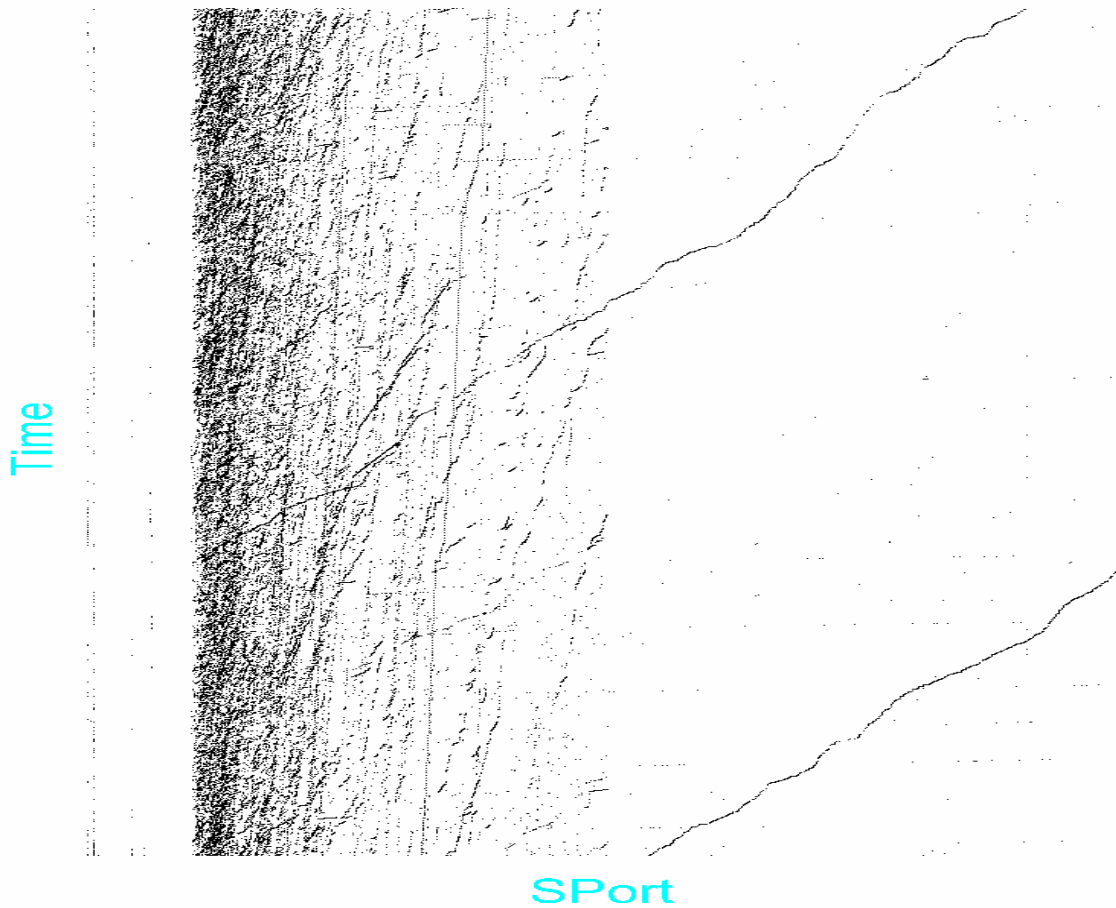
Evolutionary Graphics with
Explicit Dependence on Time

Evolutionary Graphics from Streaming Internet Data



Waterfall
for
Destination
IP versus
Time for
only one
hour

Evolutionary Graphics from Streaming Internet Data



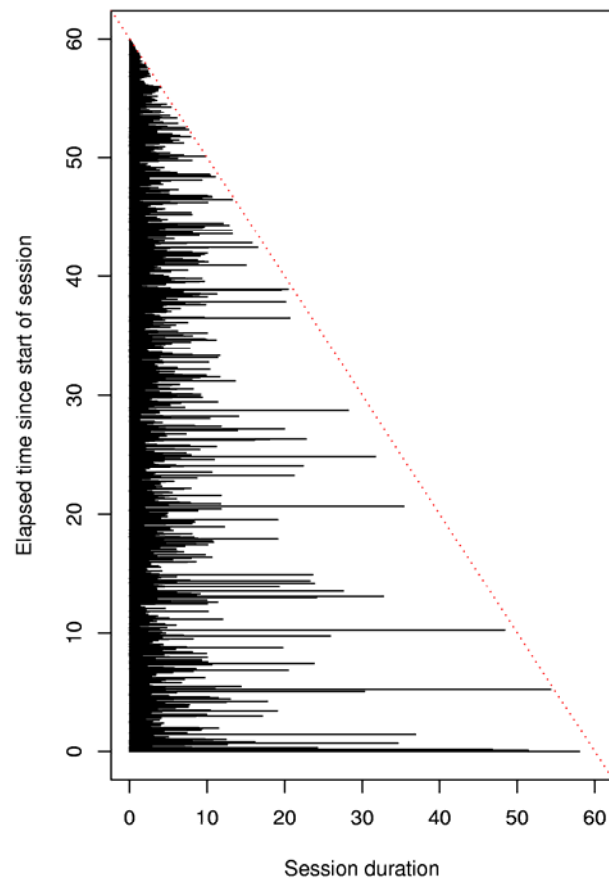
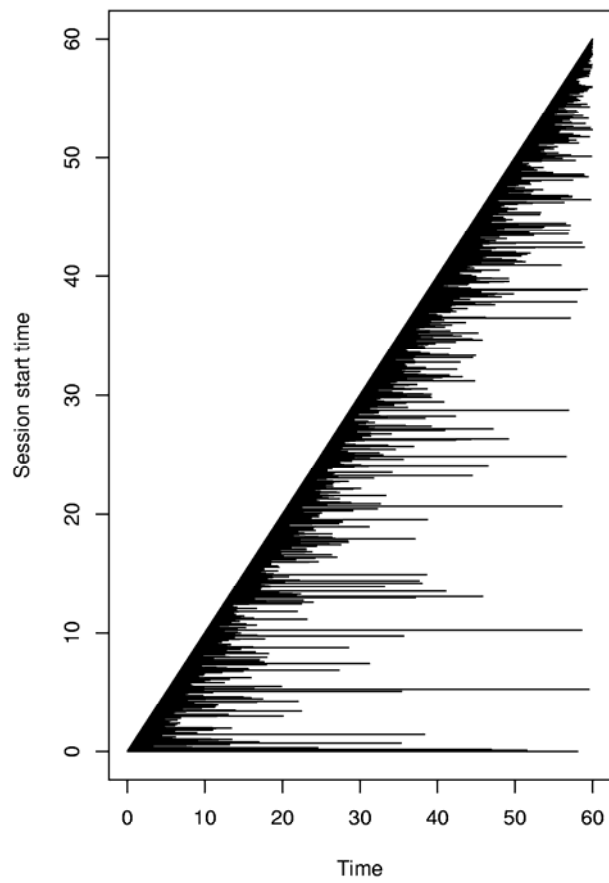
Waterfall for Source Port versus time. Diagonals are characteristics of distinct operating systems.

Evolutionary Graphics from Streaming Internet Data

The preponderance of relatively short sessions can be seen in the next figure, which displays the session durations as horizontal lines that extend from the start time to the end time. Because these data are collected in the order in which they occurred, the session start times range from time 0 (bottom line) to 59.971 (nearly the end of the hour).

The second figure shows the same information, but each line is shifted back to 0. The data are censored after one hour. With continuously monitored data, the session duration lines would continue past the censoring point. Most data are not censored because 92.3% of the sessions lasted less than 30 seconds.

Evolutionary Graphics from Streaming Internet Data



Evolutionary Graphics from Streaming Internet Data

I want to introduce two critical ideas for streaming data.

1. Block Recursion

- Instead of adjusting the statistic or the graphic by the new observation, keep a small number of observations in a moving window.
- Adjust the statistic or graphic by dropping off the effect of the oldest observation and inserting the effect of the newest.
- The graphic need not explicitly show dependence on time.

2. Visual Analytics

- Be willing to dynamically transform variables so as to exploit structure.



Evolutionary Graphics from Streaming Internet Data

Examples of Block Recursion Graphics that
do not Explicitly Show a Time Variable

Evolutionary Graphics from Streaming Internet Data

Most destination port numbers occur only once or twice during the hour; of the 380 distinct DPorts, 293 occurred only once, 47 occurred twice, 8 occurred 3 times, 5 occurred 4 times.

The remaining 27 ports occurred more than 5 times; the exceptional counts are DPort 80 (web, 116,134 times), 25 (mail-smtp, 6,186 times), 443 (secure web, 11,627), 554 (streaming video/audio, 200 times), and 113 (128 times).

Setting aside the “well-known” ports 0-1023, we plot the occurrence of destination ports numbered 1024 and above, which should arise more or less at random, and flag as unusual any Dport that is referenced more than 10 times.

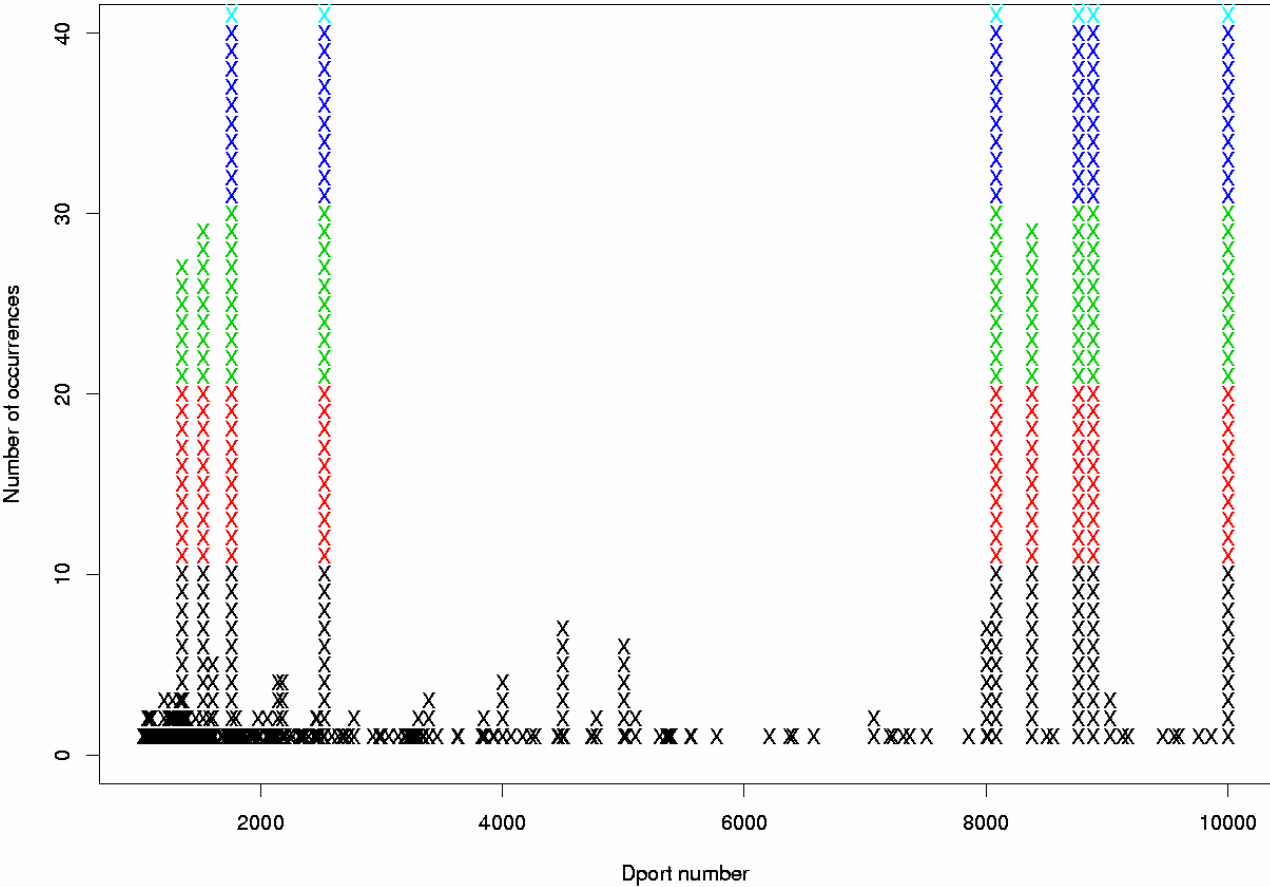
Evolutionary Graphics from Streaming Internet Data

The next figure shows such a plot; once a destination port number occurs more than 10 times, the color changes to red, indicating potentially high traffic on this destination port.

The construction of this plot resembles the tracing of a skyline, so we call it a “skyline plot.” A similar figure can be used to monitor SPort activity; however, the plot is more dense because this file contained 6,742 unique SPorts, versus only 380 distinct DPorts.

Also, while most of the 380 DPorts appeared only once in the file, a SPort typically occurred four times, with 20% of the 6742 accessed source ports occurring between 14 and 88 times.

Evolutionary Graphics from Streaming Internet Data



Animation of this plot could easily be accomplished as new data enter the block and old data depart.

Evolutionary Graphics from Streaming Internet Data

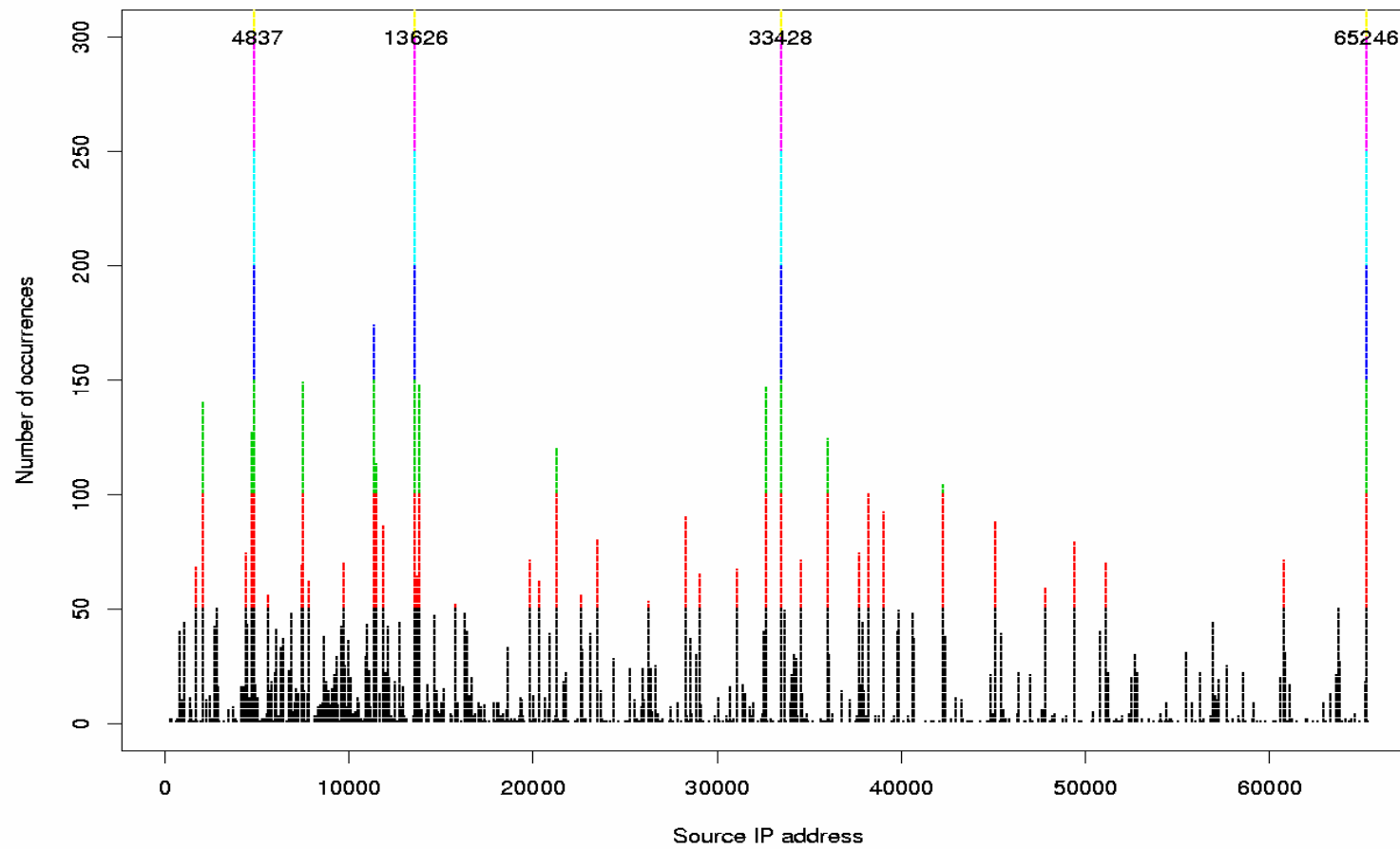
Source and destination IP addresses can be monitored also using skyline plot. In the present data file, source IP addresses are numerous (2504 unique SIPs) and frequent (the median number of occurrences is 4, and 10% occur more than 135 times).

The next figure shows this type of plot for source IP addresses in the first 10000 records, where the colors change as the number of hits exceeds multiples of 50.

Four unusually frequent source IP addresses are immediately evident in this plot: 4837, 13626, 33428, and 65246, which occur 371, 422, 479, and 926 times, respectively, in the first 10,000 sessions.

The limit for unusually frequent SIP addresses may depend upon the network and the time of day, so the limits on this “skyline” plot may need to be adjusted accordingly.

Evolutionary Graphics from Streaming Internet Data

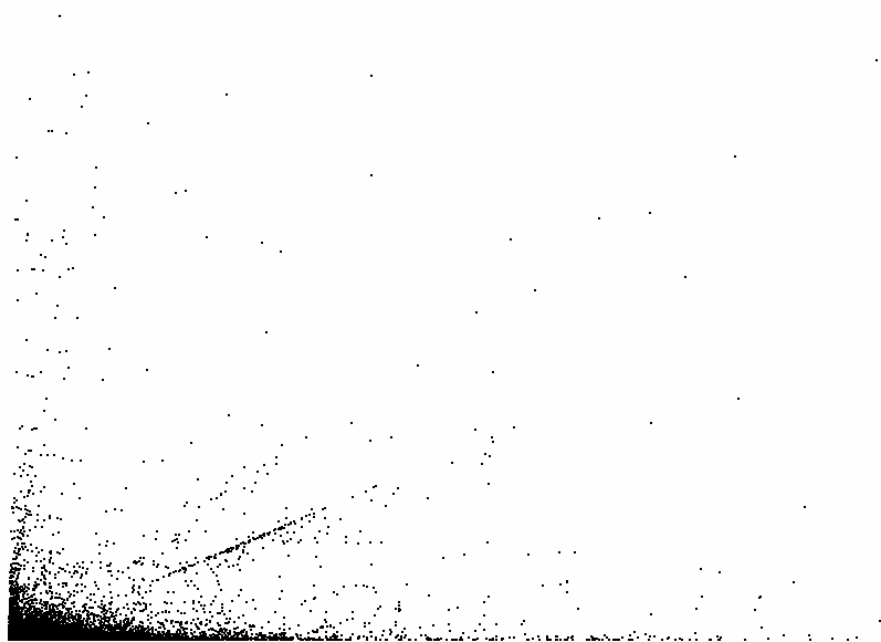


Evolutionary Graphics from Streaming Internet Data

Dynamic Transformation of Variables may be Extremely Helpful in Evolutionary Graphics. In General, for Streaming Internet Data, "Size Variables" Tend to Cluster Near Zero, so LOG and/or SQRT Transformations Are Helpful for Visual Analytics

Evolutionary Graphics from Streaming Internet Data

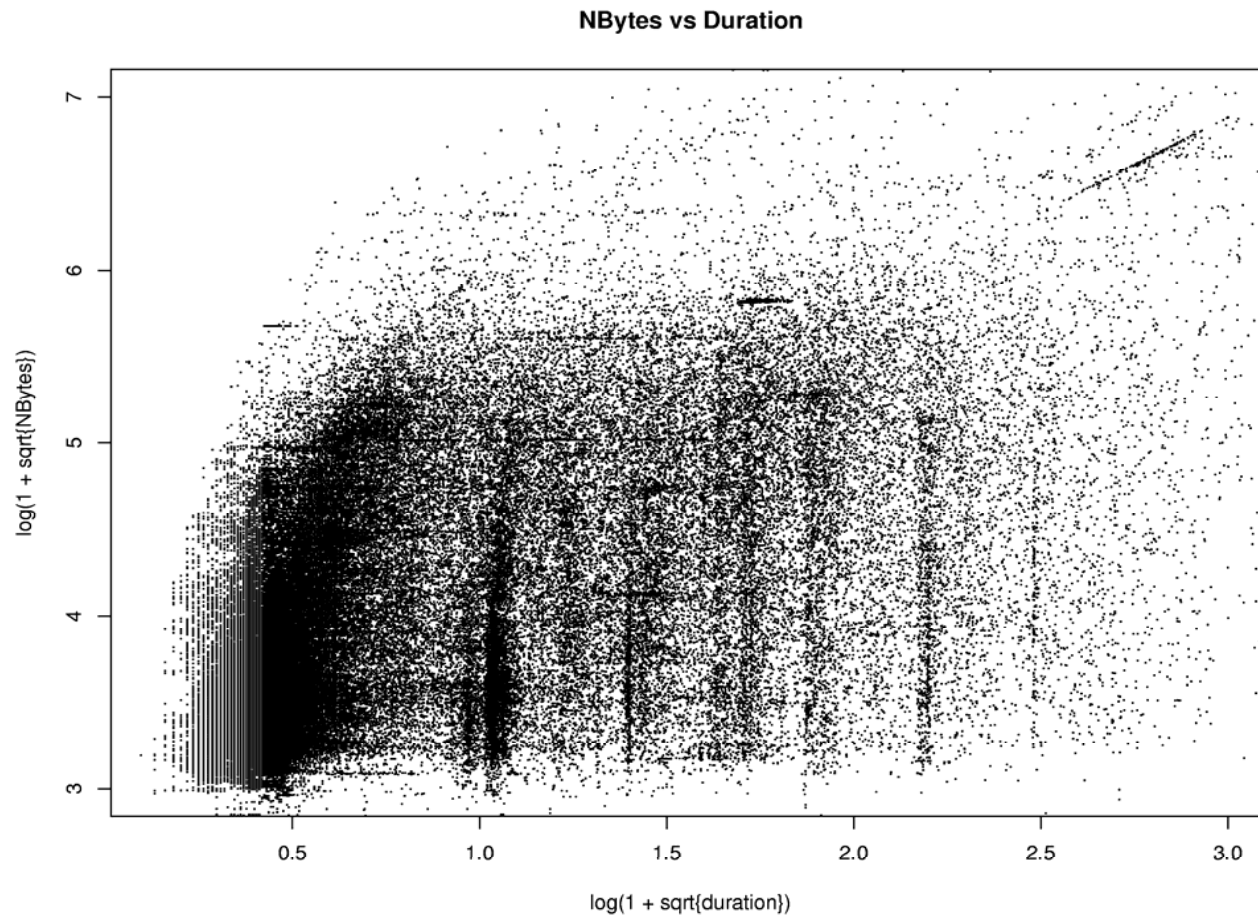
NBytes



Duration

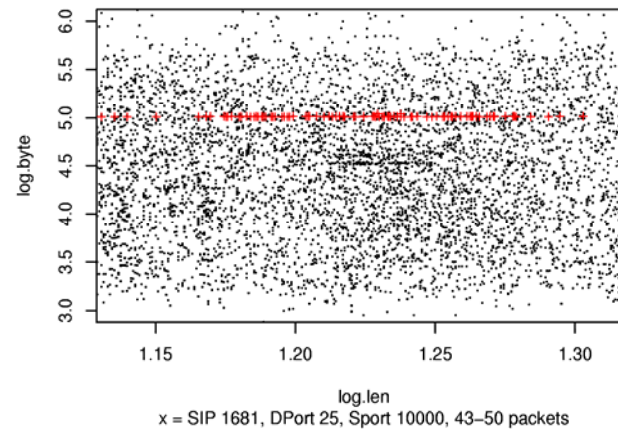
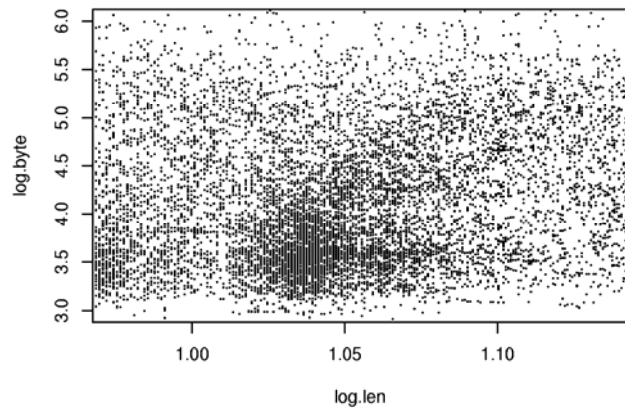
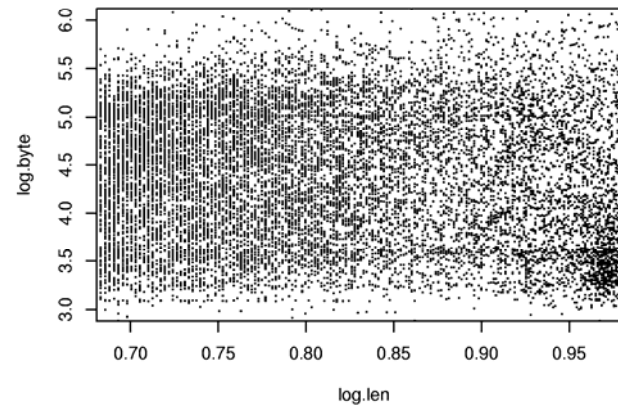
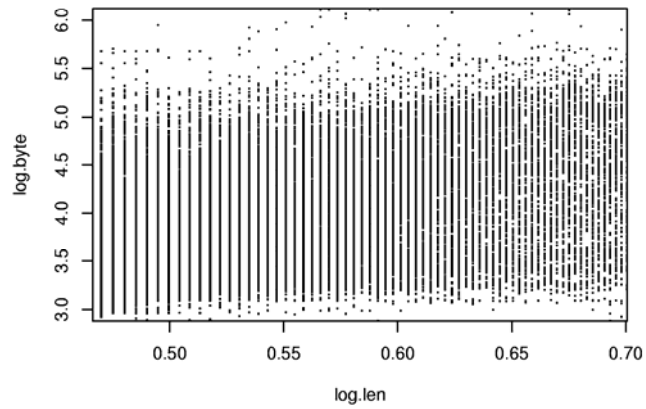
Number of Bytes versus Duration at full scale with no transformation.

Visual Analytics for Streaming Internet Traffic



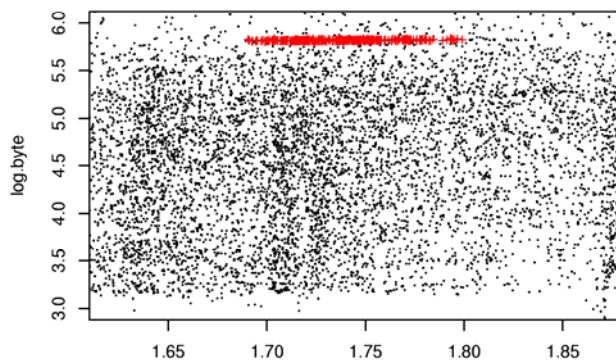
Same data as previous image with $\log(1 + \sqrt{.})$ transformation applied. Notice the additional structure visible.

Visual Analytics for Streaming Internet Traffic

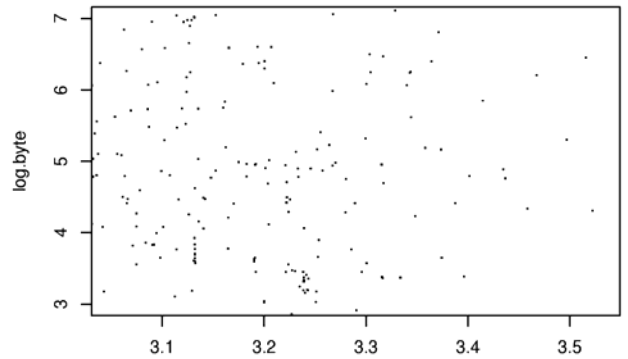
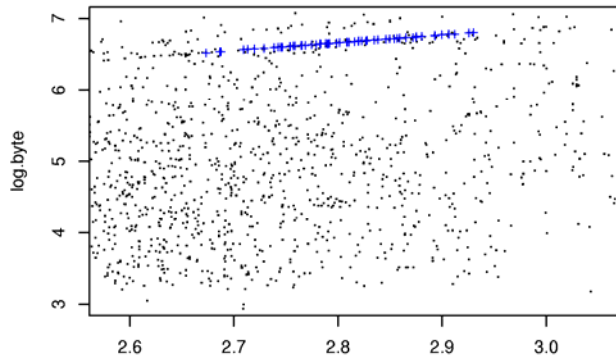
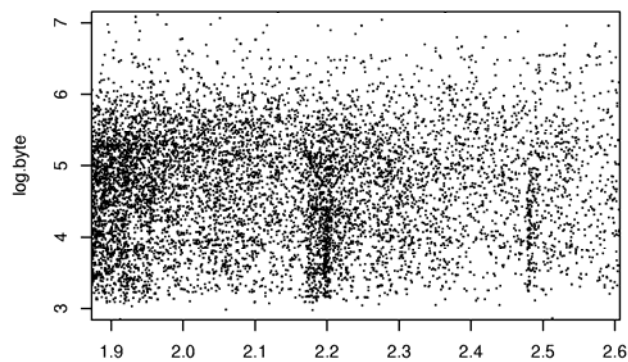


Conditional plots show additional points of interest.

Visual Analytics for Streaming Internet Traffic



292 points: SIP 23070, DIP 336, DPort 80



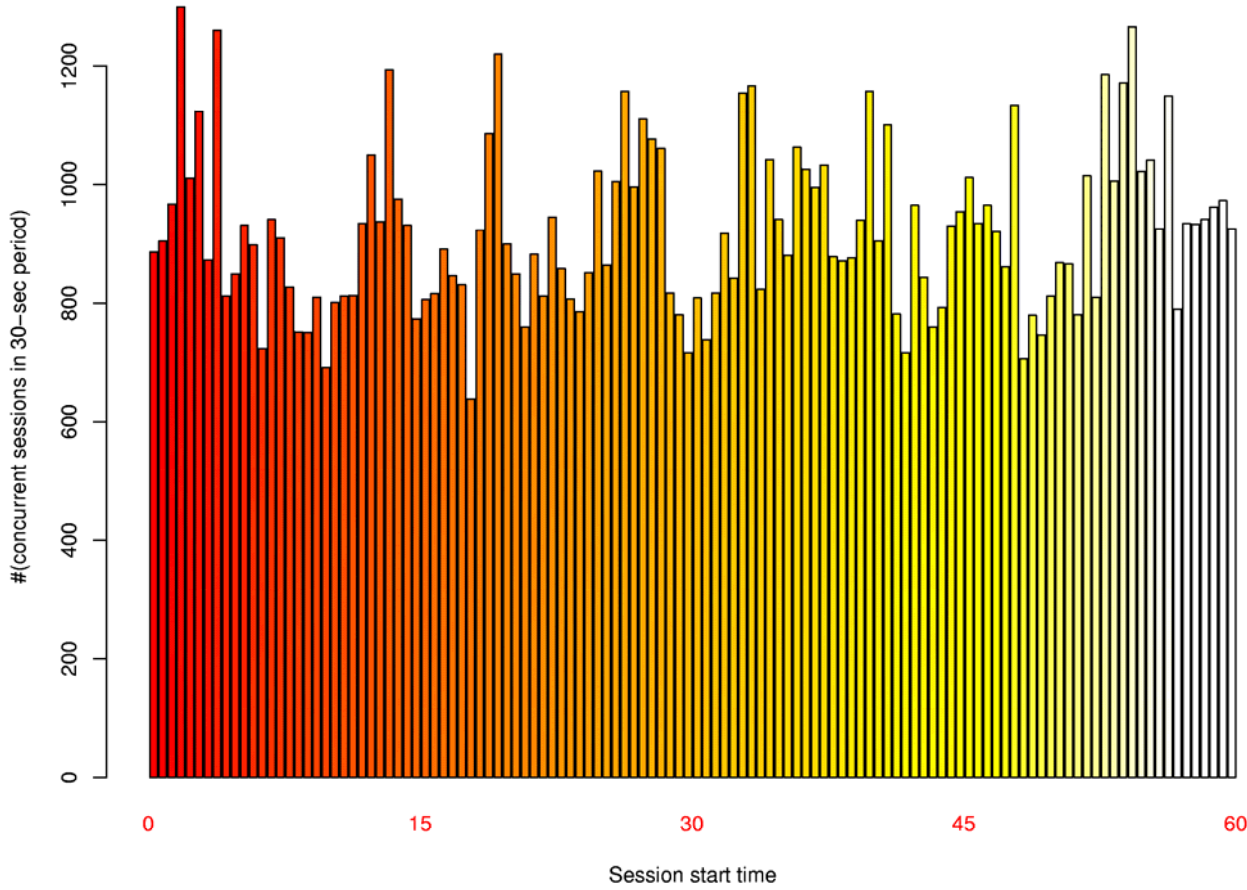
Evolutionary Graphics from Streaming Internet Data

The next figure shows a barplot of the number of active sessions during each 30-second subset of this one-hour period (a time frame of 30 seconds is selected to minimize the correlation between counts in adjacent bars).

The mean number of active sessions in any one 30-second interval during this hour is 923, and the standard deviation is about 140, suggesting an approximate upper 3-sigma limit of 1343 sessions.

A square root transformation may be appropriate. The mean and standard deviation of the square root of the counts is 30.29 and 2.23, respectively, resulting in an approximate upper 3-sigma limit of 1367, very close to the limit on the raw counts, since the Poisson distribution with a high mean resembles closely the Gaussian distribution.

Evolutionary Graphics from Streaming Internet Data



An Evolutionary Graphic

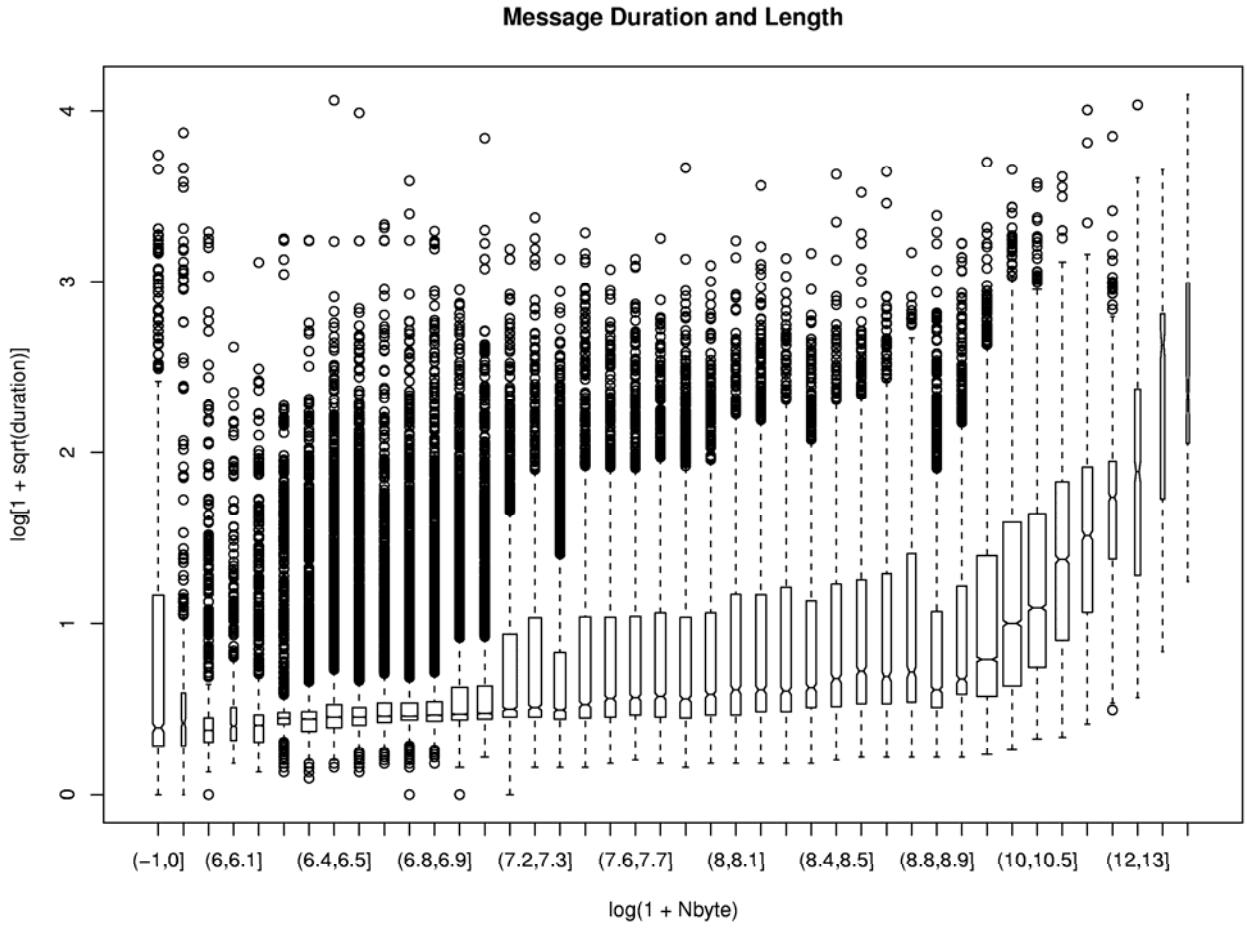
Visual Analytics for Streaming Internet Traffic

Box plots are useful for displaying the relationship between two variables. In the next plot, we plot $\log.\text{len}$ versus $\log.\text{byte}$.

The first box contains the 2611 values for which N_{byte} is zero, the next box contains 1216 values where $1 < N_{\text{byte}} \leq 365$.

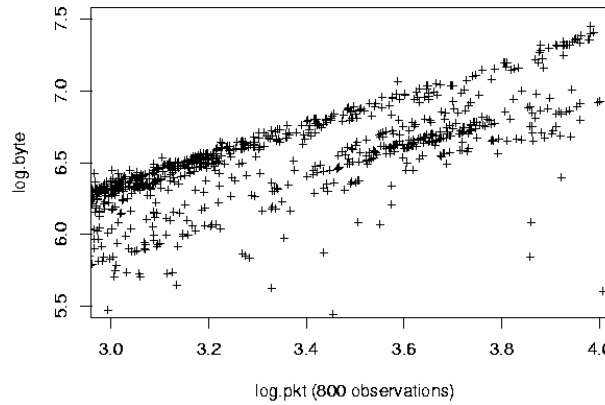
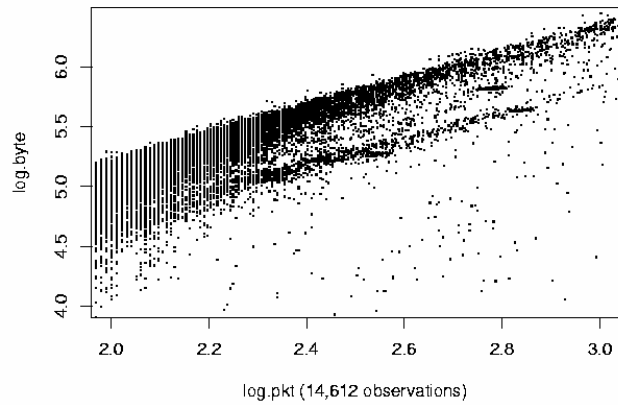
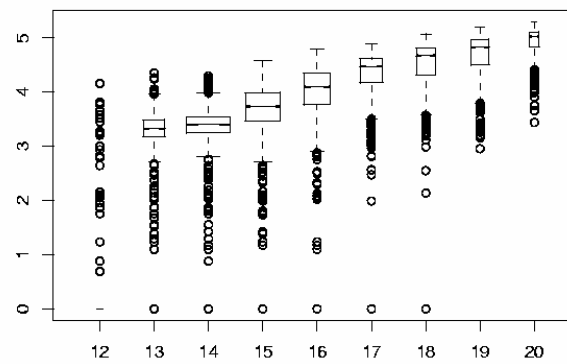
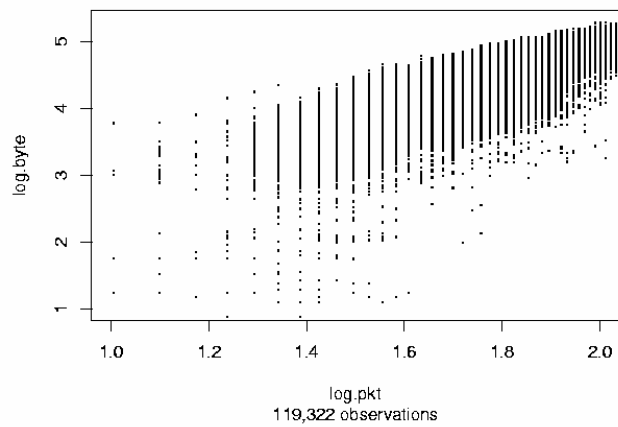
The distributions are clearly skewed to the left (a lot of outliers of large values). The distribution of duration as a function of N_{bytes} is fairly smooth and has a reasonable trend upwards.

Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

NByte vs Npacket

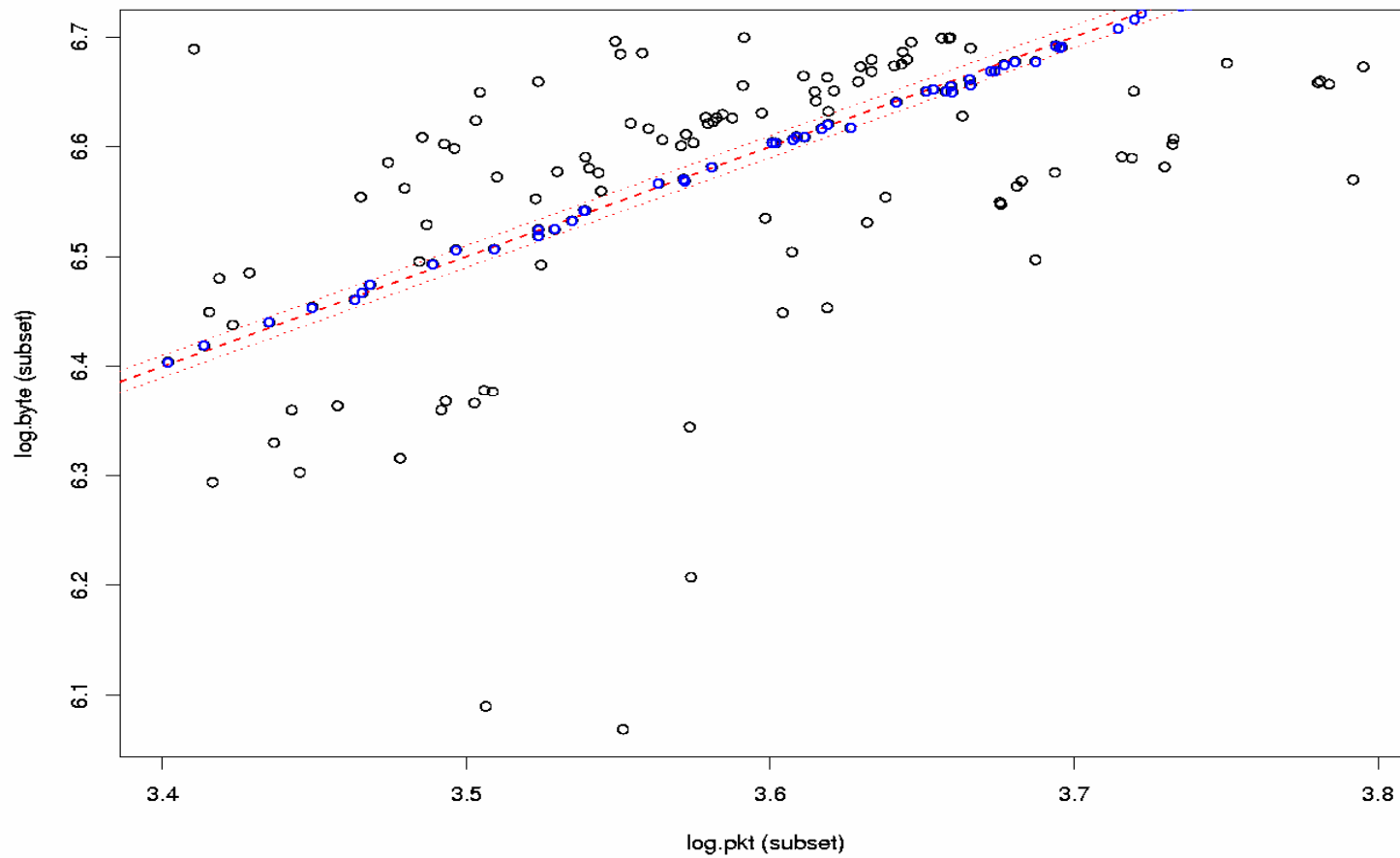


Visual Analytics for Streaming Internet Traffic

The dense set of 55 points (points in Panel (d), i.e. $3.4 < \log.\text{pkt} < 3.8$) are plotted in the next figure; they lie on or very near the line $\log.\text{byte} = 3 + \log.\text{pkt}$, and 43 of them correspond to DPort 43.

The extent to which such a pattern could occur by chance alone should be investigated, particularly if they occur all within a few seconds of each other (these did not).

Visual Analytics for Streaming Internet Traffic



Visual Analytics for Streaming Internet Traffic

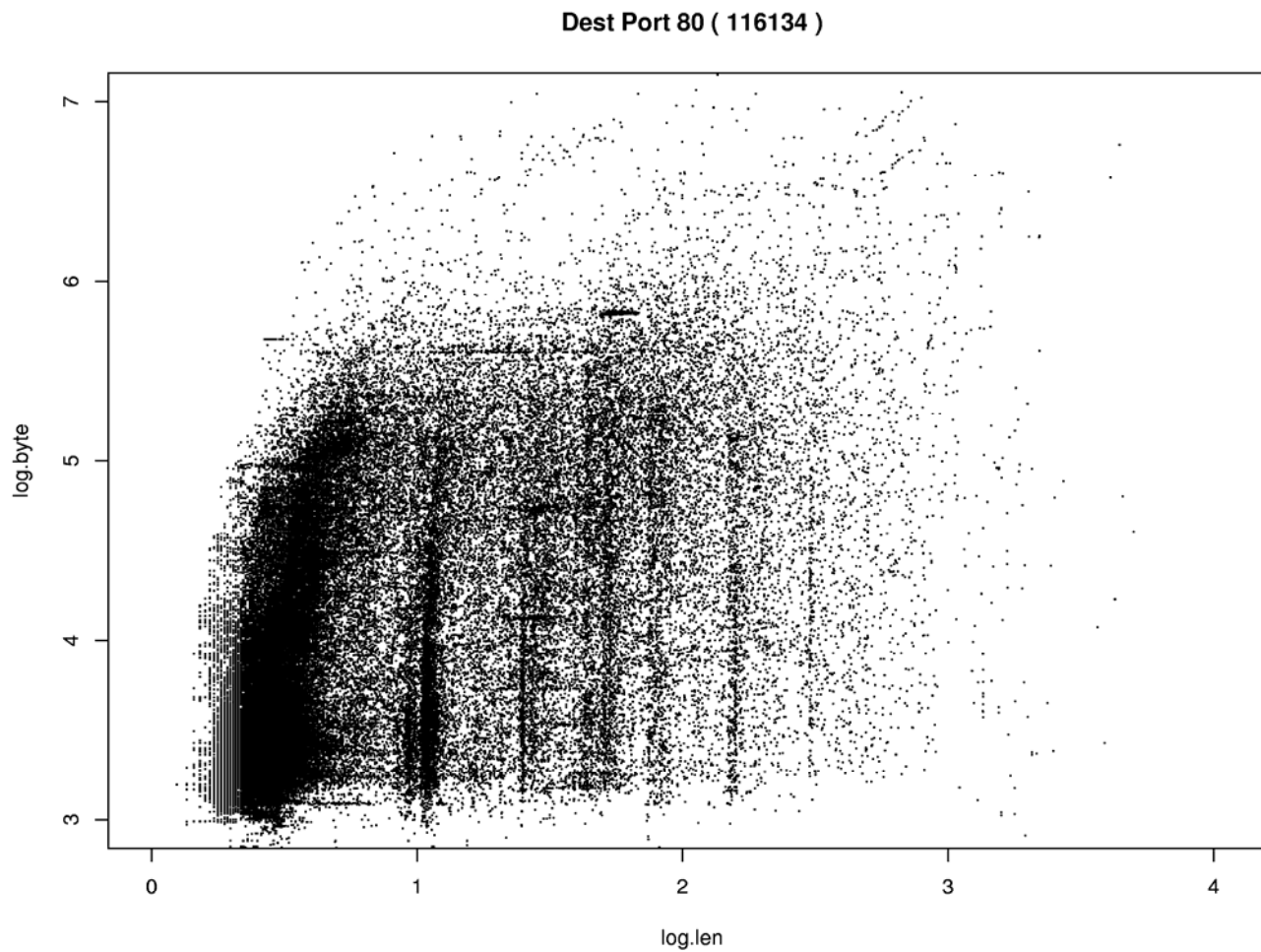
This hour of internet activity involved 380 different destination ports, DPort.

1. DPort 80 (web) is the most common, comprising 116,134 of the 135,605 records.
2. The next most common destination port is DPort 443 (secure web, https), used 11,627 times.
3. Followed by DPort 25 (mail SMTP) accessed 6,186 times.
4. Ports 554, 113, 10000, 8888 occur 200, 128, 97, 94 times, respectively.
5. Displaying all 135,605 points on one plot is not very informative, so instead we provide conditional plots according to their destination ports.

Visual Analytics for Streaming Internet Traffic

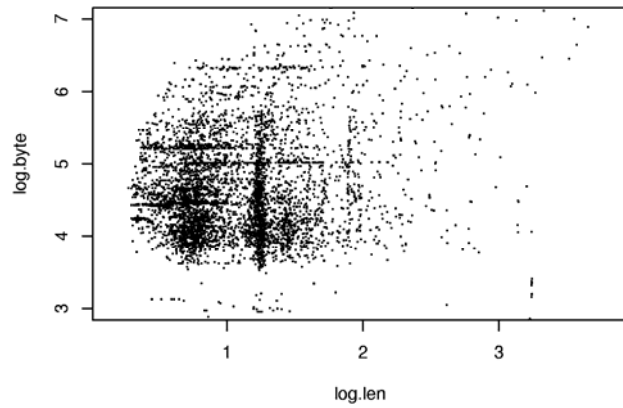
The next figure plots $\log.\text{byte}$ versus $\log.\text{len}$ for only the web sessions. A few horizontal lines of the sort noticed in previous plots appear, but otherwise no real structure is apparent.

Visual Analytics for Streaming Internet Traffic

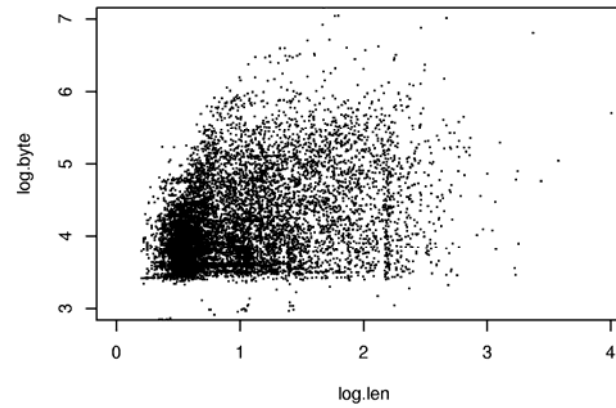


Visual Analytics for Streaming Internet Traffic

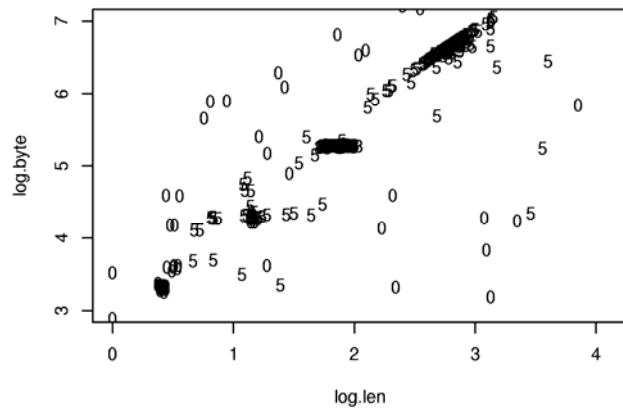
Dest Port 25 (6186)



Dest Port 443 (11627)

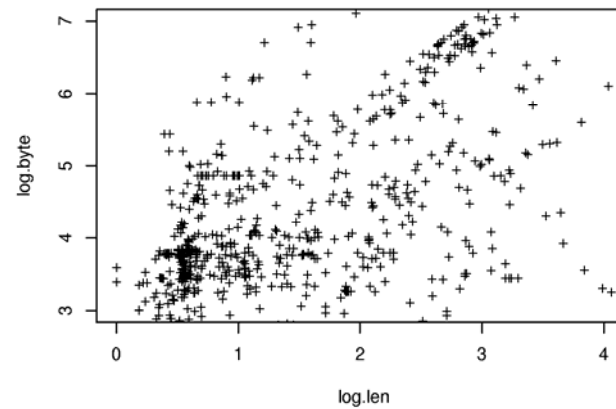


Dest Ports 113, 554, 8888, 10000 (519)



1 = 113 5 = 554 8 = 8888 0 = 1000

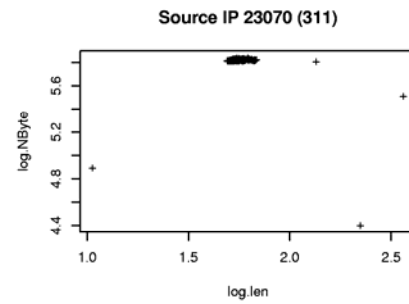
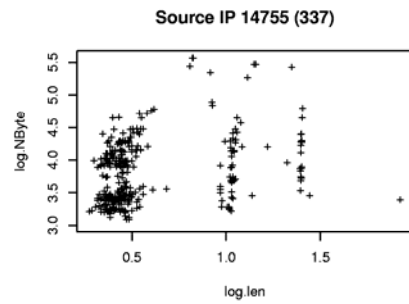
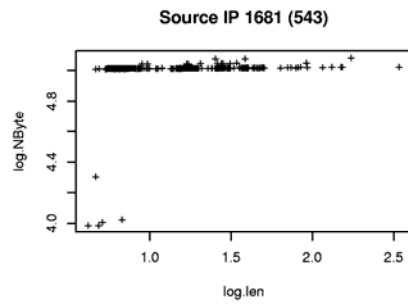
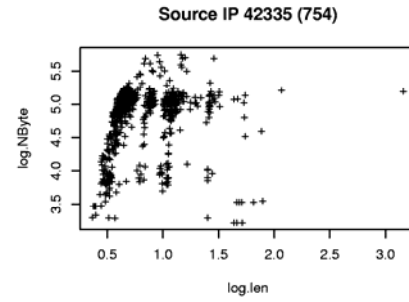
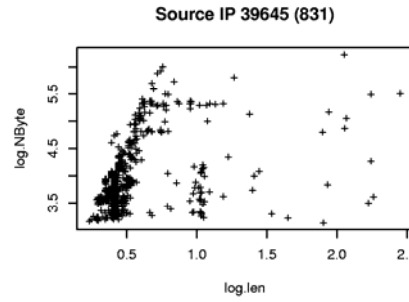
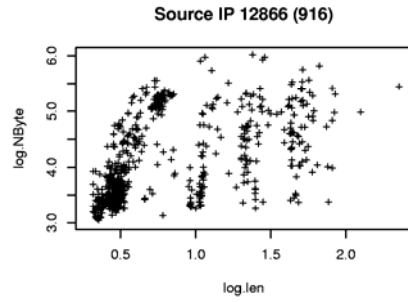
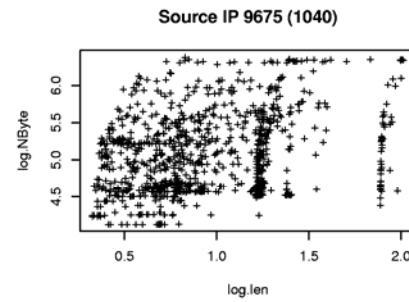
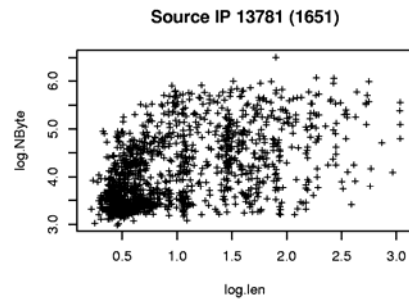
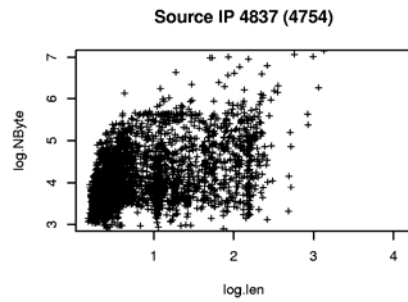
Other Dest Ports (1139)



Visual Analytics for Streaming Internet Traffic

These same plots can be constructed when the data are conditioned by source IP address, SIP, as opposed to destination port number, DPort. The number of source IP addresses that may be active during a given hour of activity is likely to be very much higher than the number of destination ports; in this data set, only 380 unique destination ports were accessed, while 3548 source IPs are in the file.

Visual Analytics for Streaming Internet Traffic

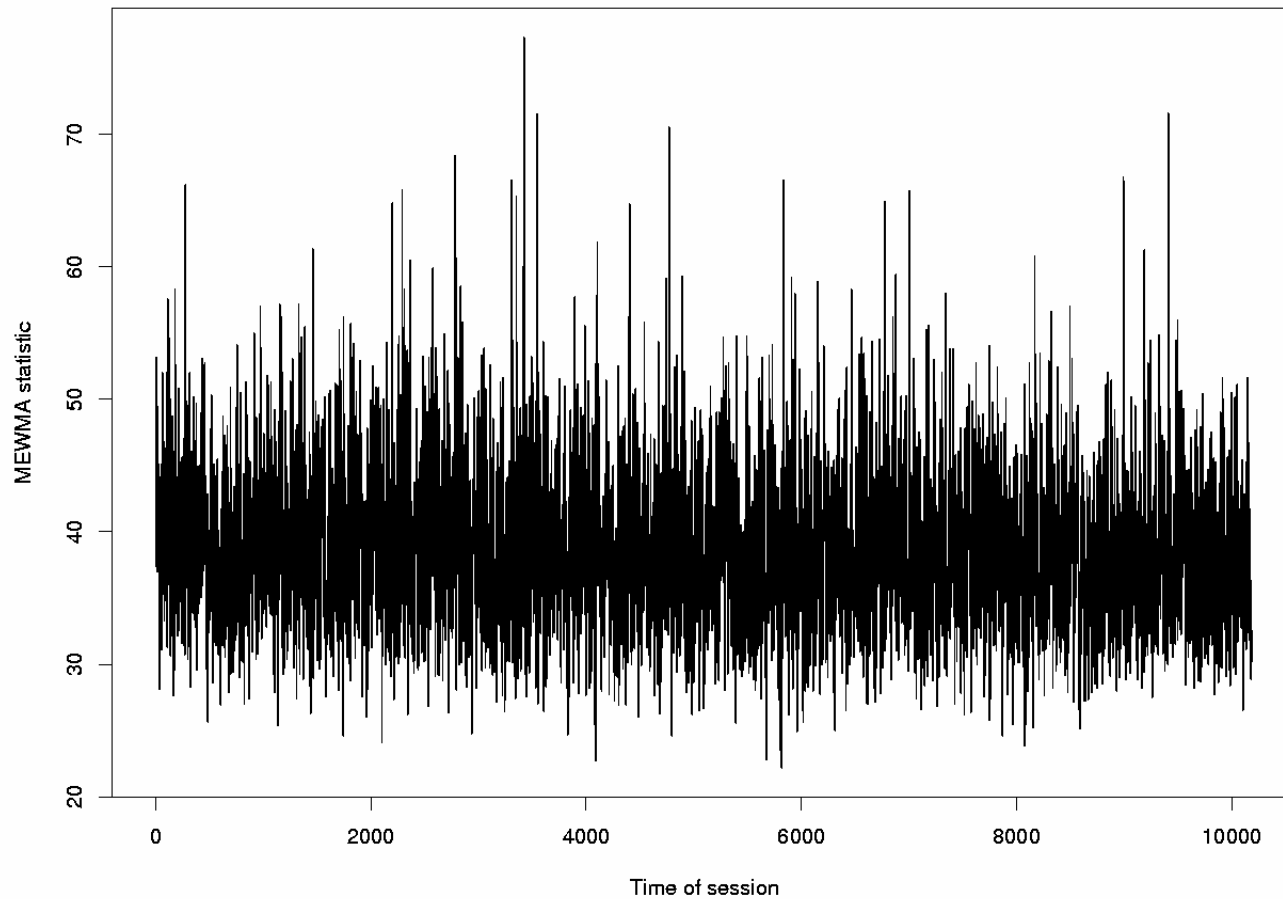


Visual Analytics for Streaming Internet Traffic

1. The three session “size” variables, $\log.\text{len}$, $\log.\text{pkt}$, $\log.\text{byte}$, are somewhat correlated and are amenable to a “control chart” procedures, where the statistic being plotted is a weighted linear combination of the previously plotted variable (λ) and the current value of Hotelling's T^2 statistic ($1 - \lambda$).
2. Calculating a Hotelling's T^2 statistic on three successive observations, denoted H_t , a multivariate exponentially weighted moving average (MEWMA) chart using $\lambda = 0.5$ is shown in the next figure (last 10,202 observations only).
3. Most values (99.7%) are below 60; a successive run of observations above 60 might suggest abnormal session sizes.

Visual Analytics for Streaming Internet Traffic

EWMA on T-Squared ($\lambda = 0.5$)



This also is a evolutionary graphic, but with a distinctly visual analytic interpretation.

Visual Analytics for Streaming Internet Traffic

Acknowledgements:

Collaborators:

Karen Kafadar, David Marchette, Jeffrey Solka, Don Faxon, John Rigsby

Research Funding:

ONR, ARO, AFOSR, NSF, DARPA at one or more stages.

Visual Analytics for Streaming Internet Traffic

Contact Information:

Edward J. Wegman
Center for Computational Statistics
George Mason University, MS 4A7
4400 University Drive
Fairfax, VA 22030-4444

Email: ewegman@galaxy.gmu.edu

Phone: (703) 993-1691

FAX: (703) 993-1700