امتحان پایان ترم اندازه گیری استاد درس: جناب دکتر خونساری

مدت امتحان: ۴۵ دقیقه

سوال اول: SNMP

الف) اجزای اصلی در SNMP چه هستند؟

ب) تفاوت SNMPv1، SNMPv2c و SNMPv3 چیست؟

ج) ساختار درختی MIB چگونه است؟

سوال دوم: کتاب Marchette

مطالب ذکر شده در ادامه چه موضوعی را بیان میکنند و از چه تکنیک هایی استفاده میکند، برای هرکدام از این تکنیک ها شرح دهید چه چیزی را نشان میدهند و چه مزایا و معایبی دارند.

2.5 VISUALIZING NETWORK TRAFFIC

We have seen a number of ways to visualize network traffic in the preceding sections. In this section, we will look at some of these techniques in a little more detail and discuss some techniques that might not be familiar to many people working in computer security.

We will start with the simplest technique, the scatter plot. This will lead us to pairs plots, which are a way to display higher-dimensional data, which will in turn lead us to parallel coordinates. These techniques will be illustrated on network data.

2.5.1 Scatter Plots

The simplest (and arguably most powerful) technique for visualization of data is the scatter plot. In this technique, bivariate data are plotted as points in a plane, with coordinates corresponding to their values. We have already seen this in previous sections (for example, Figure 2.3). Another example, Figure 2.9, depicts about a minute and a half of incoming TCP packets to a site. Time is depicted on the x-axis and destination port is on the y-axis. Note that sessions are quite

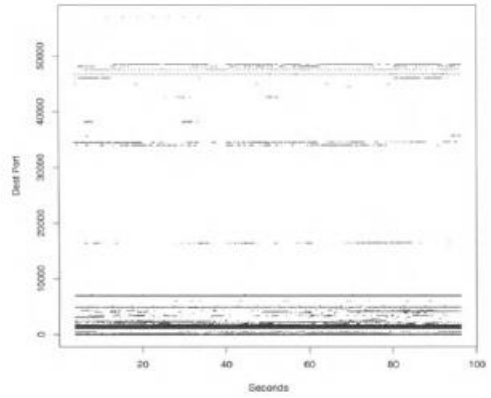

Fig. 2.9 Scatter plot of 93 seconds worth of data incoming to a site. In this case, there are 67,134 observations. The destination port number is plotted against time.

clear in this depiction as horizontal line segments. We can also see banding effects corresponding to port ranges for popular applications.

It is instructive to zoom in on this plot. Figure 2.10 depicts the data for which the destination port is less than 10,000, which is where the bulk of the data lie in Figure 2.9. Here we see an interesting phenomenon in the range between 40 and 80

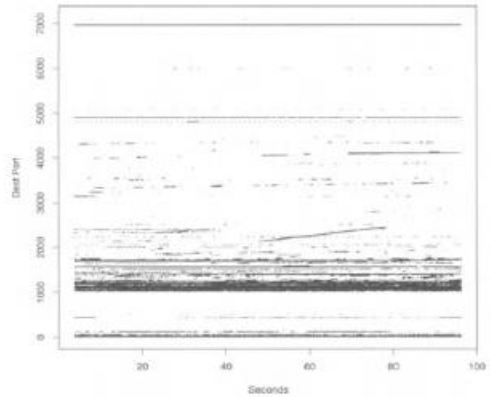

Fig. 2.10 Scatter plot of packets from Figure 2.9 with destination port number less than 10,000.

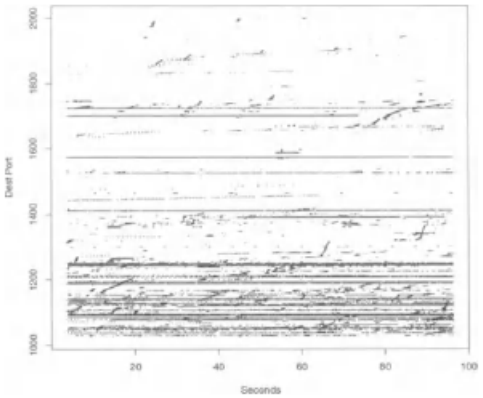

Fig. 2.11 Scatter plot of packets from Figure 2.9 with destination port between 1000 and 2000.

seconds. There is a line that is angled slightly, corresponding to destination ports between 2000 and 2500. Some further investigation shows this to be a Web session (the source port is 80). In this case, one of the site's users has gone to a Web server, and the downloads from the Web site are happening on the higher-numbered ports (which is typical Web client behavior).

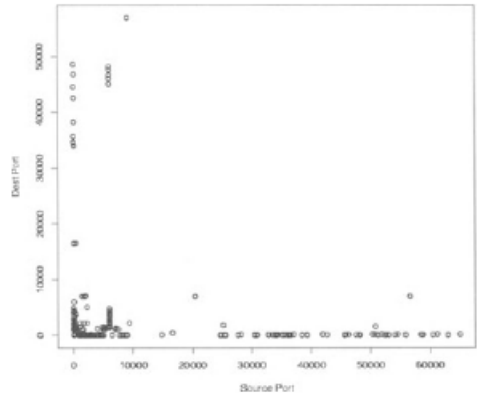

Fig. 2.12 Scatter plot of source port against destination port for the data from Figure 2.9.

This plot does not allow us to see that the Web session described above is actually a session to a single machine. In other words, although the line in the figure appears to indicate a single session we cannot tell that these packets are in fact related. They may be going to several different machines and the apparent correlation may simply be a coincidence. This is unlikely, especially given the length of the line but we cannot rule it out from this plot.

One way to solve this problem would be to use color to encode the destination (or source) machine. This would allow us to pick out sessions much more easily, but there are limitations to the number of colors that humans can reliably distinguish. Also, there can be problems with overplotting, especially when attempting to depict large time intervals for networks with heavy traffic loads. This requires interactive exploration where the user chooses a number of different views of the data; for example, zooming in to regions of interest.

A further zoom of these data is depicted in Figure 2.11. Here we can see quite a bit of fine structure. There are several angled lines, indicating data transfers of some kind, as well as a lot of horizontal lines, indicating application sessions on a number of ports. Again, color could be used to enhance this picture. Further zooms can also be used to explore different regions in the data.

Another way to look at these data is to plot source port against destination port. This is done in Figure 2.12. The "L" shape of this plot is indicative of the tendency for applications to use low-order ports. Thus, one can usually infer the application that corresponds to the packets as the one corresponding to the smaller of the two ports.

This is not a particularly useful plot. However, there are some outliers in the plot and by definition outliers are interesting. These outliers stand out quite easily in this plot, while they would be very hard to detect by looking in the raw data. A good rule of thumb is to look at any data via several different kinds of plots. Look for "interesting" structure and for "abnormal" or unusual data. The definition of unusual is subjective, but like art, one usually can recognize unusual data when one sees them.

Scatter plots are arguably one of the most powerful ways to visualize data, primarily because of their simplicity of interpretations. Their main drawback is the inherent low-dimensionality of the data that can easily be displayed in a scatter plot. In the rest of this section we will look at ways to plot higher-dimensional information.

Another drawback to scatter plots, which is common to all plots, is the problem of overplotting. There are only so many pixels on the screen, or dots on the paper, and so there are only so many distinct dots that can be represented. This problem can be partially addressed by binning the data and depicting the bins as is done with histograms, or by interactive displays which allow the user to zoom in to areas of high density. Unfortunately, paper generally does not lend itself well to user interaction.

2.5.2 Pairs Plots

An obvious extension of a scatter plot to higher-dimensional data is to plot each pair of variates in a separate scatter plot. This is the idea behind a pairs plot. We

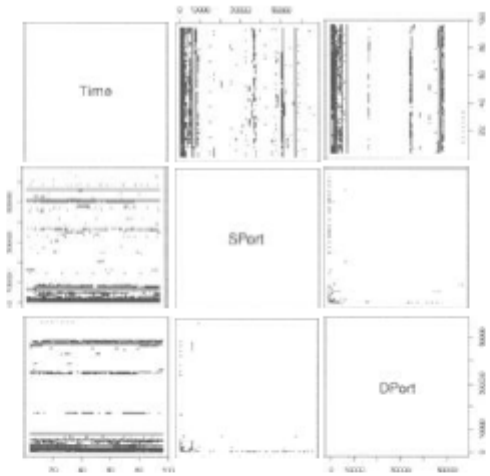

Fig. 2.13 Pairs plot 90 seconds of packets to a site. SPort and DPort represent the source and destination ports, respectively.

reconsider the data discussed in Figures 2.9–2.12, plotted as a pairs plot in Figure 2.13. The pairs plot is symmetric about the diagonal, so we are actually plotting twice as many plots as we need. However, sometimes it helps to see things from two perspectives, so I will use the default plot of the R pairs function in these plots.

We can see from these plots that the low-value destination ports span both time and source port, whereas the high destination ports are only associated with relatively low source ports. This corresponds to our earlier analysis. What cannot

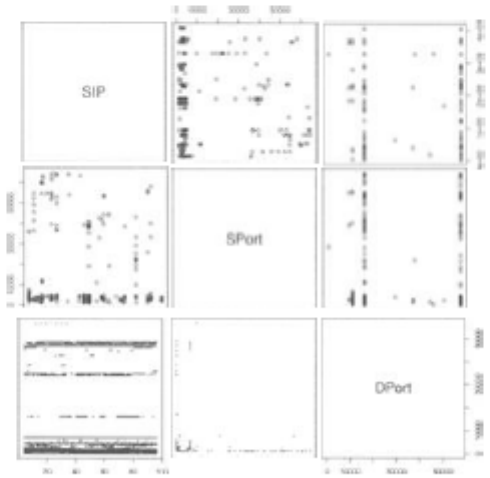

Fig. 2.14 Pairs plot of email (port 25) connections. SIP and DIP represent the source and destination IP addresses, while SPort is the source port. In this case, all the destination ports are port 25.

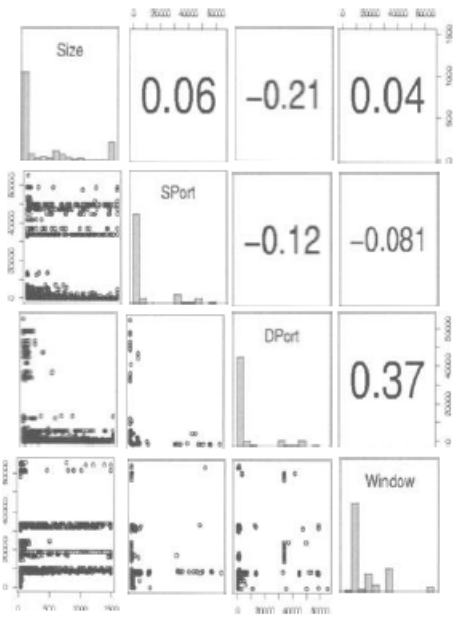

Fig. 2.15 Pairs plot depicting the packet size, source and destination ports and window size for 13,680 TCP packets. A histogram of each variate is plotted down the diagonal, and the correlations are given in the upper triangle.

be determined with single scatter plots is the combined information of all three variables. For example, consider the largest destination port in the SPort by DPort plot. Looking at this value in the Dport by time plot, we see a very regular pattern (points spaced roughly 3.5 seconds apart). This turns out to be activity between source port 9100 and destination port 56,946. This is an interesting pattern. Further investigation of the data determined that the session was initiated by the monitored site, and the temporal pattern was probably a result of load at one end or a delay in the route (other data between these two machines did not show the 3.5 second delay between packets).

Figure 2.14 depicts data to port 25 (email) for one hour's worth of data. The three variables source IP, source port, and destination IP are plotted (source and destination IP having been converted to 32-bit numbers). For example, the plot in the upper right corner has vertical axis source IP and horizontal axis destination IP. From this, we can count roughly nine destination IPS (the resolution of the plot makes this an inexact count). We can see that most sources send to one of two destinations (presumably the main mail servers for the site). From the SIP by SPort plot, we can see that most machines use source ports below 10,000, although some use higher ports.

Personally, I am not particularly impressed with pairs plots, particularly for more than three variables. I find that I have trouble visually processing more than three plots at a time, and so tend to not use pairs plots for much of my data analysis. Occasionally, though, they do provide useful information and so are a part of my visualization toolbox.

The pairs plots depicted here have a lot of redundancy and unused space. This extra plotting area could be used much more efficiently. One use of the extra plotting area would be to utilize color. One could color the points according to a separate scheme in the upper triangle of the pairs plot than in the lower. Another use would be to provide a different zoom level. The R function "pairs" allows separate functions of the data to be plotted in the different panels.

One interesting idea is to put a histogram of each variate down the diagonal of the plot. This is illustrated in Figure 2.15. In this plot the histogram for each variable is displayed in the diagonal. Also, in the upper triangle are the correlations for the pairs. Thus, we can see that the window size is most correlated with the destination port while packet size is inversely correlated with destination port. These correlations are fairly low. Note that source port is uncorrelated with window size, as is packet size. This illustrates the fact that the window size is a feature of the application, and has nothing to do with the size of the packet.

Pairs plots can usefully display up to about a dozen variables, depending on the amount and complexity of the data to be displayed. Other methods are required for higher-dimensional displays. We will look at two such methods in the next sections.