Functional Analysis of Computer Network Data

J. L. Solka
Code B10
NSWCDD
Dahlgren, VA 22448

D. J. Marchette
Code B10
NSWCDD
Dahlgren, VA 22448

Abstract

This paper examines the application of hierarchical cluster analysis to the characterization of single machine SYN ACK time series. The purpose of the analysis is the identification of normal and abnormal activity for a small group of mail and web servers. Ultimately this information could be used to help analyze forensic data or as part of an on-line network or host based intrusion detection system.

Keywords

time series analysis, hierarchical cluster analysis, network traffic analysis

1 Introduction

Computer services at universities or in the private or governmental sectors almost always include email and web servers. These machine are typically involved in activity with other machines on a 24/7 time frame. One can think of the hourly patterns of activity as a signature of the traffic associated with a particular day. Given this mind-set it makes sense to ask the question which daily activity records are similar and which differ. Once this question is answered then one can proceed forward with an investigation into the mechanism that generated the similarity or lack thereof among the days.

We confine our analysis for the most part to the examination of the activity records associated with a particular machine. Although this seemed like the best approach for our initial study, one could expand the effort to include the study of daily activity logs for several machines offering the same service. For example one could examine hourly time series for all of the multiple web servers at a particular site. This of course would complicate the analysis when one was faced with the task of ascertaining why usage patterns were the same for two different machines on a particular day or for multiple days for various machines. In fact some of our preliminary results seem to suggest that it might make sense to examine data sets obtained from groups of machines that are engaged in entirely different services. For example one might wish to utilize traffic from web and email servers. On some platforms the act of reading email and the act of surfing the web are carried out using the same software, i. e. a web browser and hence the traffic associated with the two activities may be correlated.

In the background section of the paper, we provide the necessary rudimentary introduction to network protocols and hierarchical cluster analysis so that the reader

is well poised to follow the analysis presented in the results section. This is followed by the conclusions section.

2 Background

In this section we provide some of the rudimentary background information on computer network traffic, hierarchical cluster analysis, and principal component analysis. Our intent in each of these sections is to provide the bare essentials needed in order to understand the results presented in the results section of the paper.

2.1 Network Traffic (The nature of the beast)

The transmission control protocol/internet protocol (TCP/IP) was originally developed by the Defense Advanced Research Projects Agency (DARPA) to provide a means for Department of Defense computers to communicate with one another in the advent of war. The internet is currently used to link educational, governmental, and commercial sites throughout the world.

Information that is transmitted between computers is referred to as a datagram. The IP layer of the transmission process contains routing information and control information that is associated with the delivery of the datagram. Information such as source IP address and destination IP address are contained in the IP frame header. An address is a 32 bit indicator of the location where the packet originated and is traveling to. It is of the form A.B.C.D where each of A, B, C, and D is constrained to be between 0 and 255. There is other information contained within this header, but it is not the subject of our current consideration.

TCP provides a reliable stream delivery and virtual connection service. Reliability is maintained through the use of sequenced acknowledgment with retransmission of packets when necessary. The information contained in the TCP header includes the source and destination port for the packet that is being transmitted. These ports are associated with specific applications running on the machines.

A TCP/IP connection is established in a three-way handshake procedure. First a client machine sends a packet to the server with the synchronize (syn) flag set in the TCP header section. This SYN flag is a clue to the server machine to begin the process of packet number sequencing. This process allows the two machines, client and server, to keep track of packets that may arrive in a somewhat asynchronous manner. The server then responds with a packet with a SYN and and ACK (acknowledge) flags set. The client then responds with a packet with an ACK flag set and the process proceeds forward until both the client and the server decide to terminate the connection by sending a packet with the finish (FIN) flag set.

A port is a way for the computer to use port numbers to distinguish between (demultiplex) different logical channels on the same network interface on the same computer. There are 65,536 different ports associated with modern ethernet cards. The port numbers less than 1024 are usually reserved for certain priveledged processes. These processes in a UNIX/LINUX environment are owned by root. An individual computer has helper processes, called daemons, that wait in order to "facilitate" communication on the various ports via the various protocols. The two that are of interest to us are port 25 that is usually associated with mail and port 80 that is associated with web traffic.

Depending on the setup/configuration of a machine one can ascertain whether a particular service is being run on a machine by looking for SYN/ACK packets

from the port associated with that particular service. This may not be true in some circumstances, such as certain kinds of proxy servers but certainly a server that is running a particular application will be responding with SYN/ACKS on the appropriate port.

2.2 Hierarchical Cluster Analysis

This section provides a modicum of material on agglomerative cluster analysis. The reader is referred to [Everitt, 1993] for a more thorough treatment of this material. The purpose of cluster analysis is to divide a group of data into a set of subsets where each subset possesses some sort of cohesive structure. This definition does of course leave much room for interpretation and in fact the science of cluster analysis is really more of an art. We will assume that the reader is familiar with some of the basics of the field in order to press on with our explanation.

Given a set of observations in p-dimensional space one must first develop a method to measure the distance between not only the observations but also the clusters that are formed during the clustering procedure. There are a number of ways to measure the distance between two clusters including average distance, closest distance, and most distance distant. The method using most distant distance has been employed within. This method is known as complete clustering.The results obtained within were obtained using the agglomerative hierarchical clustering procedure. In this procedure one begins with a single cluster at each of the observations and then at each stage of the process one merges the two closest clusters.

The agglomerative procedure produces a range of possible clusterings of a given data set all of the way from as many clusters as observations to a single cluster. the user typically desires to have a single answer for the number of clusters. This decision can be made based on a visual inspection of the dendrogram or can be made in an automated fashion. A dendrogram is really just a way to present to the user a visual portrayal of the agglomerative clustering procedure. It is effectively a tree where clusters are plotted as they are merged during the clustering procedure.

Alternatively one can attempt to ascertain the number of clusters in an automated fashion. The choice of the number of clusters is a subject of continued research within the statistics community. This is due in part to the difficult nature of the process. In fact under the most general assumptions it is probably impossible to ascertain the "true" number of clusters and their inherent structure. The method of Mojena is used to ascertain the number of clusters in the port activity time series data [Mojena, 1977].

Mojena suggested that one should select the number of groups corresponding to the stage in the dendrogram where

$$\alpha_{j+1} > \overline{\alpha} + k s_{\alpha} \tag{1}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to the stages with $n, n-1, \ldots, 1$ clusters. The terms $\overline{\alpha}$ and s_{alpha} are defined as the mean and unbiased standard deviation of the α-values and k is a constant. Mojena recommends that k should be chosen between 2.75 and 3.50. Milligan and Cooper [Milligan and Cooper, 1985] suggest a value of k of 1.25 and this is the value that is used in the results section.

2.3 Principal Component Analysis

Principle component analysis was performed using the R prcomp command. This command performs a principal component analysis on a data matrix via a singular value decomposition on the centered data matrix. The data matrix is usually transformed to have a unit variance in each of the variables. In our case we have chosen to perform the principal component analysis on the centered unscaled data. We felt that scaling was unnecessary since each hours counts was approximately the same order of magnitude when averaged across the multitude of days.

3 Results

Our study was conducted using traffic collected during the time period of March 1, 2000 till August 31, 2000. Data was obtained using the TCPDUMP utility on a number of machines at our site. These data were subsequently parsed using PERL. All of the statistical analysis was performed using the public domain statistics package R.

We have chosen to characterize a machine's activity on a particular port as a time series consisting of SYN/ACK packet counts for each of the hours from midnight on a given day till the following midnight on the next day. This level of time granularity was chosen as a compromise between a finer granularity of minutes or a coarser granularity of days. Ultimately the purpose is to consider these time series as points in a twenty four dimensional space and then to subsequently subject these points to cluster analysis.

We will focus our discussions within on three of our sites machines. One machine is an external mail server and we will be examining its SYN/ACK time series counts on port 25. The second machine is an external web server and we will study its SYN/ACK time series counts on port 80. The final machine is an internal web server and we will also examine its traffic on port 80. In Figure 1 we present a plot of the $\log_{10}$(SYN/ACK counts) vs. time for the external mail server port 25 on the left and the external web server port 80 on the right. We notice the similarity of the primary component of the curves for the two plots. This is probably caused by the diurnal cycle of business here at our site. Most people tend to arrive at work around 8:00 am and depart around 5:00 pm. This time of arrival/departure schedule is probably responsible for the shape of the primary component of the two curves. The zero values corresponds to times with no data, usually because of the sensor or machine being off-line.

It is interesting to compare Figure 1 to a similar plot, Figure 2 (port 80), for the internal web server. Since the internal web server was not involved in nearly the level of traffic as the two external servers one does not see the characteristic usage pattern that is present in the previous plot. We also call the reader's attention to the large traffic levels associated with the early morning hours. These are indicated by the high peaks on the two ends of the plot. We will have a little more to say about these peaks later.

Let's now turn our attention to a closer analysis of the external web server traffic counts. In Figure 3 we present a plot of the dendrogram obtained via a hierarchical agglomerative cluster analysis of the port 80 counts associated with this machine. The complete method was used to measure the distance between the clusters during the agglomerative procedure. This method measures the distance between two clusters as the distance between the two most distant observations in the two clusters. The dendrogram indicates several single point clusters, a cluster of

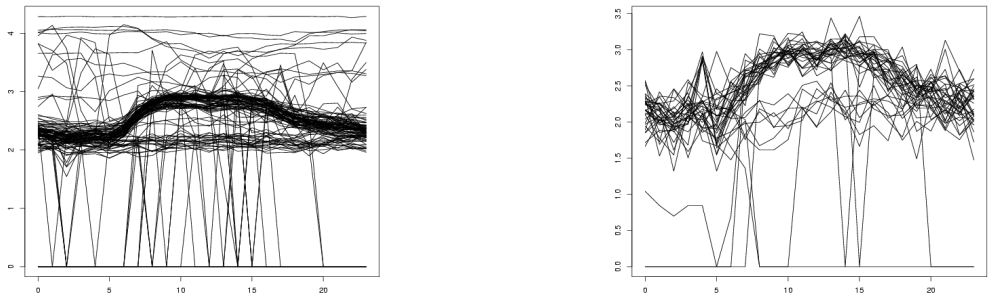

Figure 1: Representative traffic curves for the external mail server, port 25, and the external web server, port 80. Time period is March 1, 2000 to August 31, 2000.

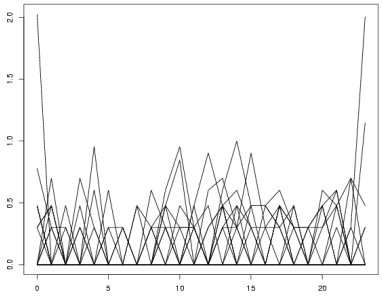

Figure 2: Representative traffic curves for an internal web server, port 80. Time period is March 1, 2000 to August 31,2000.

3 days:July 19, August 19, and August 23, along with a a larger cluster, indicated in the center of the dendrogram, and a cluster of 7 days on the right hand side of the dendrogram. The 3 day cluster of July 19, August 19, and August 23 will be the subject of later discussions.

We next turn our attention to the examination of the mean count curves that are associated with several of the external web server clusters. In Figure 4 we present a plot of the mean profiles for clusters 1, 2, 3, and 4. Cluster 1, plotted in the upper left hand corner, consists of 31 days of relatively constant traffic levels. The second plot in the upper right hand corner corresponds to August 14. There is a curious artifact in this plot in that all of the counts reach a level of 0 around 8:00 am in the morning. It is difficult to track down the cause of such an artifact in the data during a post mortem analysis. This type of drop out could be caused by the web server itself going down or by the sensor that is collected the data for our analysis going down. Often careful records are not kept as to which of these two situations occurred. The plot in the lower left hand corner indicates the activities associated with August 18. Once again we do see some curious valleys. The final plot in the lower right hand corner presents a mean activity plot for July 19, August 19, and August 23. This is the 3 day cluster that was mentioned in the previous dendrogram discussion section. The differences in these clusters appears to be mostly the resul

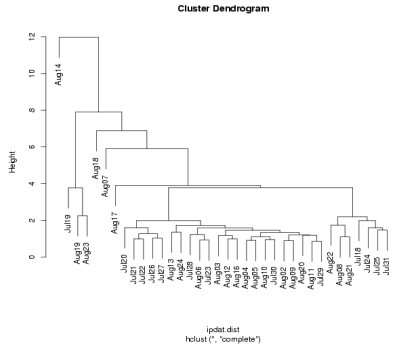

Figure 3: External web server dendrogram obtained using complete agglomerative based hierarchical clustering. Time period is March 1, 2000 to August 31,2000.

of the "drop-outs."

Next we examine the cluster structure associated with the external mail server. In Figure 5 we present the dendrogram associated with the agglomerative hierarchical cluster analysis of the external mail server. Once again we have used the complete method of measuring distances between clusters. This particular dendrogram is plagued by poor over plotting due to a standard layout procedure. Its layout could be improved by the use of hyperbolic coordinates.

We can overcome this problem by projecting the observations into the first two principle components. In Figure 6 we present a plot of the external mail server clusters rendered in the first two principal components. We draw the reader's attention to a few of the clusters that are resident in the plot. The cluster labeled 8 consists of the single day June 14. The cluster labeled with the symbol 1 consists of the days April 1, 10, 17, 27; May 12; June 12, 26; July 19; August 19. The cluster labeled with the symbol 6 consists of the days May 10, June 17, and August 14. The cluster labeled 1 is of particular interest to our follow on analysis.

It is illuminating to compared the mean count time series between the external web server and the external mail server. In Figure 7 we present mean count time series for several of the web server and mail server clusters. We have positioned the web server mean count time series on the left hand side of the page and the mail server time series on the right hand side of the page. The upper left hand plot consists of the days July 19, August 19, and August 23. The upper right hand plot consists of the days April 1, 10, 17, 27; May 12; June 12, 26; July 19; and August 19. It is important to note that there was no web server traffic/data recorded for April 1, 10, 17, 27; May 12; and June 12 and 26. With this revelation in mind, one can note a high degree of overlap between the cluster structure for the web server portrayed in the upper left hand plot and the traffic for the mail server portrayed in the upper right hand figure. In fact one can even ascertain a similar character to the mean count time series for these two clusters. Turning our attention to the bottom two plots we notice that these two both share the day of August 14. An examination of there mean count time series seem to suggest that there was a sensor/network failure on August 14. This is evidenced by the fact that both of the time series went to 0 at around 8:00 am.

The last set of profiles to consider may be the most interesting. Consider the mean activity profiles presented in Figure 8. These represent single point clusters

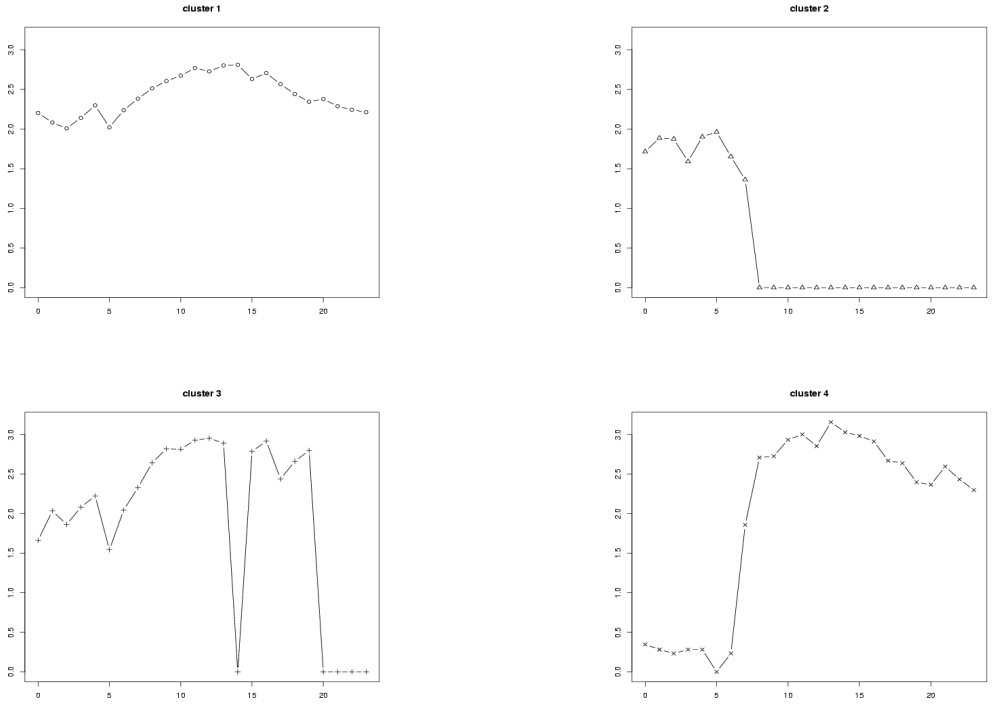

Figure 4: Mean count cluster profiles in external web server traffic. (a) 31 day cluster (b) August 14th Monday. (c) August 18th Friday. (d) August 19th Saturday, August 23rd Wednesday, and July 19th Wednesday.

obtained from a hierarchical cluster analysis of the internal web server traffic. The plot on the left corresponds to August 23 and the plot on the right corresponds to August 24. Before proceeding with our analysis of these plots let us make a couple of casual observations about the rest of the clusters associated with this machine/port combination. There is a large cluster of days associated with no traffic, several single point clusters, and these two "extreme value " clusters. The two anomalous days August 23 and August 24 are a curiosity. The extreme values associated with hour 23 of August 23 and hour 0 of August 24 are very curious and are the subject of our continued investigations.

4 Summary and Conclusions

We have demonstrated the use of agglomerative hierarchical cluster analysis to characterize SYN/ACK time series associated with email and web traffic on a small number of machines. This sort of analysis is complicated by transient machine, sensor, and network outages. We do feel however that this type of approach is particularly relevant to forensic traffic analysis in order to ascertain anomalous activity. This approach may potentially be useful as part of a host-based system that can detect unusual machine utilization based on port activity time series.

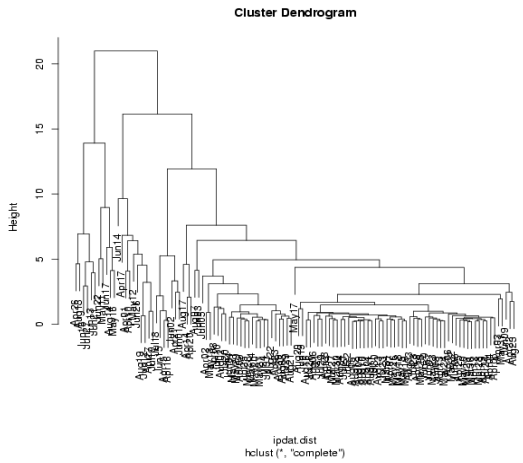

Figure 5: External mail server dendrogram obtained using complete agglomerative based hierarchical clustering. Time period is March 1, 2000 to August 31,2000.

Acknowledgements

The authors would like to thank the Office of Naval Research Code 311 for sponsoring this work.

References

[Everitt, 1993] Everitt, B. S. (1993). *Cluster Analysis*. John Wiley and Sons, New York, third edition.

[Milligan and Cooper, 1985] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *PPsychometrika*, 50:159–179.

[Mojena, 1977] Mojena, R. (1977). Hierarchical grouping methods and stoppings rules an evlauation. *Computer Journal*, 20:359–363.

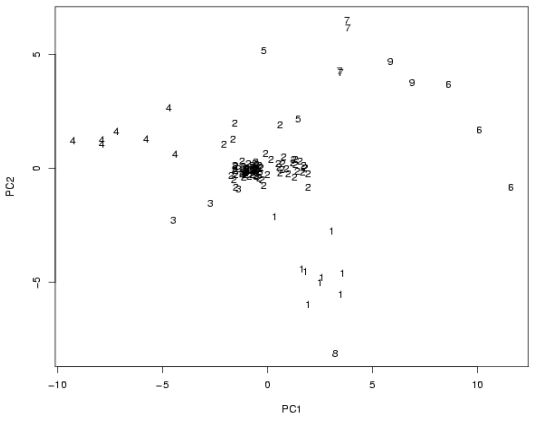

Figure 6: External mail server clusters rendered in the first two principal components.

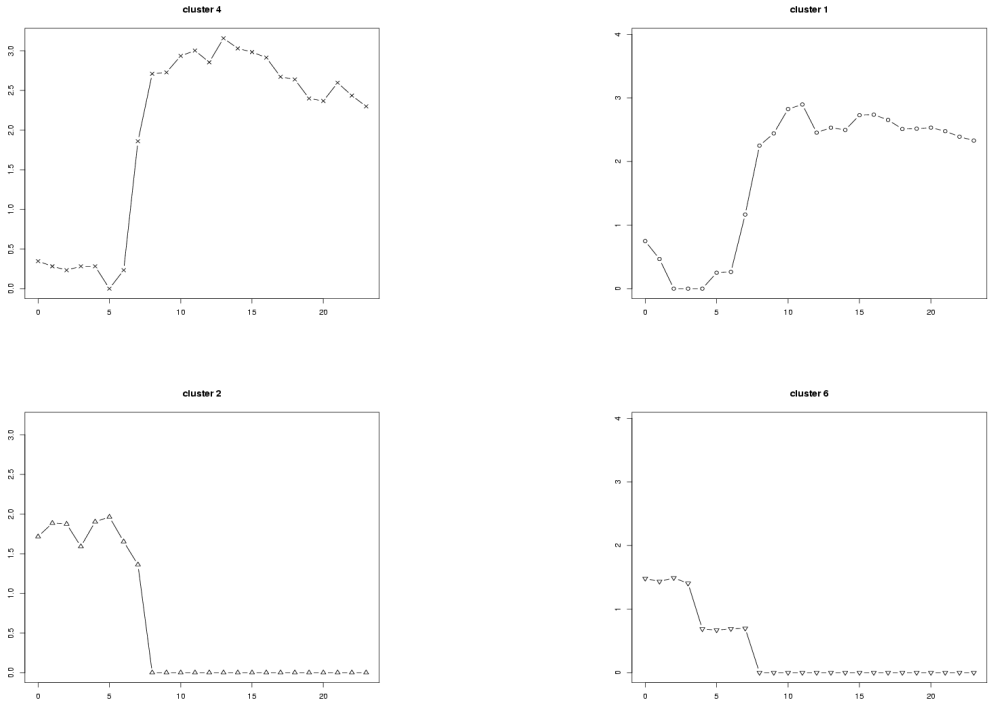

Figure 7: Comparison of mean count cluster profiles in external web server and mail server traffic. (a) July 10th, August 19th, and August 23rd. (b) April first, 10th, 17th, 27; May 12th; June 12th, 26th; July 19th; August 19 (c) August 14th. (d) May 16th; June 17th; and August 14th.

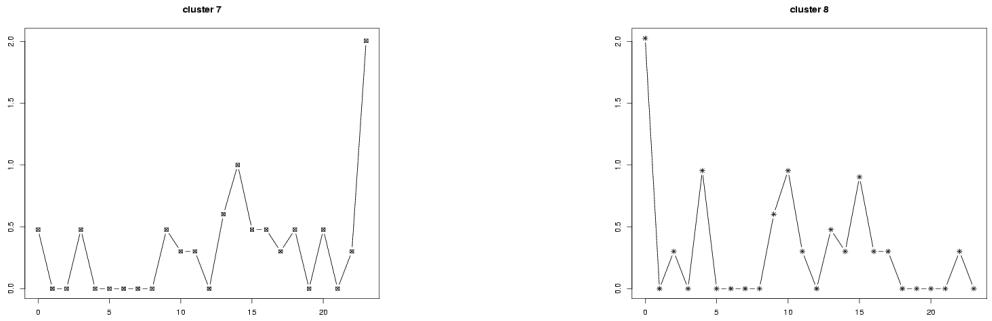

Figure 8: Interesting traffic structure for two consecutive days on the internal web server data.