

Man vs. Machine - A Study of the Ability of Statistical Methodologies to Discern Human Generated ssh Traffic from Machine Generated scp Traffic

J. L. Solka¹, M. L. Adams², and E. J. Wegman³

Abstract

This paper discusses our recent results in the classification of human-based ssh traffic as compared to machine-based scp traffic. Since both of these services, ssh and scp, use the same port, port 22, this classification problem occurs within a quite natural framework. Results that illustrate an exploratory analysis of the data will be presented along with some preliminary classification results.

Key Words: ssh, scp, pattern recognition.

1 Introduction

The attacks of September 11th have demonstrated to the United States of America and to the world the vulnerability of a country's infrastructure to attack. A country's infrastructure includes the traditional items such as transportation, communication, and energy facilities. These types of traditional infrastructure play an important role in a country, however a more important role may be played by the computer network infrastructure of the country.

These cyber infrastructure attacks are perpetuated by a variety of individuals. These individuals run the gambit from bored teenagers to people who are under the control of organization which are only loosely coupled to state such as the Al Qaeda and can even include attacks that are sponsored by specific hostile countries.

Some of these attacks require human intervention while many of the attacks specifically utilize automated tools in order to ascertain the vulnerabilities and attack particular systems. We as defenders of these systems would like to be able to tell whether the activity on a system has been under the control of a human being or machine. This paper attempts to take a small step toward this capability by exploring the capability to discern between secure shell (ssh) traffic and secure copy (scp) initiated traffic. In the first case the data stream produced during the interaction is totally under the control of the human operator. In the second case the data transference is initially human initiated but is then ultimately under the control of the computer.

The first section of this paper provides a little additional discussion as to the problem at hand. The next section examines some of the statistical problems of interest that are resident within the ssh vs. scp problem. The next section provides some of the results that we have obtained during our initial

¹ J. L. Solka is a Senior Scientist, Advanced Computation Technology Division (Code B10), 17320 Dahlgren Rd., Dahlgren, Virginia, 22448-5100 (Email: solkajl@nswc.navy.mil)

² M. L. Adams is Scientist, Advanced Computation Technology Division (Code B10), 17320 Dahlgren Rd., Dahlgren, Virginia 22448-5100 (Email: asamsml@nswc.navy.mil)

³ E. J. Wegman is the Bernard J. Dunn Professor of Information Technology and Applied Statistics, Center for Computational Statistics, George Mason University, MS 4A7, Fairfax, VA 22030-4444 (Email: ewegman@gmu.edu)

analysis of the problem and the final section besides summarizing provides a brief discussion of our future plans for continued research in this area.

2 Background

2.1 Cyber Infrastructure as a Growing Concern

The computer network infrastructure has become ingrained in many if not all of the other infrastructure components. Many of the power plant systems for example are under the control of supervisory control and data acquisition (SCADA) systems. These SCADA systems provide a clear-cut avenue for cyberspace compromises to propagate into our everyday world with potentially dire consequences.

The banking sectors have also become dependent on the successful functioning of our cyber infrastructure. Cleverly planned attacks could disrupt the delicate flows of capability within the banking sector. Finally we note that the commercial sector in general has become dependent on the unfettered flow of information. Previous distributed denial of service attacks have demonstrated the extreme vulnerability of the on-line commercial sector to attacks which seek to deny customers access to their websites.

2.2 Statistical Problems of Interest in the Cyber Infrastructure Arena

There are numerous problems of interest to the statistical community that are resident within the infrastructure protection area. Some of the relevant statistical techniques include exploratory data analysis, clustering, discriminant analysis, and visualization. The focus of our analysis within is in the area of exploratory data analysis and visualization with an eye toward discriminant analysis.

2.3 The ssh vs. scp Problem

Constraints on the length of this paper prevent us from providing a complete discussion of the inner workings of the ssh or scp service. The reader is referred to Stevens (1994) for a wonderful treatment as to the intricacies of Internet packet-based communications. scp and ssh are two services that can be configured to run on a platform in a server or client mode. ssh is essentially a telnet type program where the data stream is subject to an encryption process. Similarly scp is a file transfer or ftp type of program where the file transfer stream is also subjected to an encryption process. Typically one machine is set up as a ssh/scp server and then a group of other machines, clients, run a client program that allows them to access the ssh or scp capability of the server. In this manner the client can login to the server and run a typical terminal session or even tunnel various application through the ssh "pipe".

The part of this process that is relevant to our discussions is the fact that we are attempting our analysis not on each of the single packets that are exchanged during the session. Each session or interaction between two computers are based on the exchange of a sequence of packets. We have chosen from the onset of our analysis to work with sessions rather than individual packets. It would be virtually impossible to discern the difference between a single ssh and scp packet.

The ssh and scp services are typically configured so that they both run on port 22 of the server machine. In this manner packets associated with either scp or ssh sessions on the server machine would both be transmitted on an identical port. This fact is the focus of our analysis in that we are interested in rather one can distinguish between ssh and scp traffic on this port.

3 Results

Here we describe our experimental protocol for data collection/processing, provide our results, and discuss them.

3.1 Data Collection Process

We initially planned to collect data using a server running scp and ssh on port 22 and then distinguish among the two services using user provided log files. It is also important to note that our facility really supports a single user per machine and hence we also have the capability to distinguish a particular user based on their IP address. This approach proved untenable in that each of the particular machines on our network presents a slightly different clock time due to different rates of central processing unit (cpu) clock drift on the system.

We addressed this data collection problem by modifying the configuration of the server in order to run ssh on port 76 and scp on port 77. Each user participating in the study then aliased their scp and ssh commands so that they accessed these non-standard ports on our server. In this way we could trivially disambiguate the scp and ssh traffic.

3.2 Session Tracking Process

As mentioned in the previous section we do not use the raw packet signatures in our scp/ssh recognition scheme but rather use information/statistics computed on sessions. A session for our purposes can loosely be defined as a set of packet exchanges between a user on a particular client machine and our pre-configured server from initiation of the connection to the final termination of the session.

We used an in-house developed package designated as TRACKER or XTRACKER to take a sequence of packets associated with a particular session and assemble them into the associated session so that we can compute various statistics on the session.

3.3 Session Statistics Calculation

The features extracted for each session are as follows: day of the week, year, month, date, seconds since the beginning of time, source IP address, source port, destination IP address, destination port, number of packets transmitted during the session, number of packets into the server, number of packets out from the server, number of data packets (non-header packets) transmitted, number of data bytes transmitted, duration of the session, and status of the session. A few minor clarifications about these features are in order. First the beginning of time is measured from January 1, 1970, second the source port is the port on the client machine that the session was initiated from (usually a random high numbered port), the destination port is of course 76 or 77

depending on the service being requested, header packets are those portion of the datagram that contain some of the non-data type information associated with the packet, and status of the session is a flag that indicates the final status of the session. The status of the session flag is currently not employed during our analysis. The remainder of the features was chosen based on the current capabilities of the TRACKER/XTRACKER programs and our intuition regarding which features would be more fruitful in our quest to distinguish between scp and ssh traffic.

3.4 Visual Discernment of scp/ssh Class Structure

In this section we examine some of the various features that we have examined in the scp vs. ssh discriminant analysis problem. The first two-dimensional feature pair are *data packets* and *total packets*. The reasoning behind this feature pair is that the ratio of data packets to total packets for scp sessions should be greater than the ratio of data packets to total packets for the ssh sessions. In Figure 1, we plot number of session data packets vs. number of total packets for the ssh and scp sessions. The ssh sessions are labeled by red plus signs while the scp sessions are labeled by blue circles. A visual assessment of this plot would suggest that there would be little discriminant utility in this particular feature pair. Visual assessment of feature utility is a very tenuous process and we have chosen to also evaluate each feature pair using a simple classification scheme. We have chosen to use a one-nearest neighbor classifier and quantified the classifier performance using a standard leave one out cross validation scheme. The reader is referred to Duda, Hart, and Stork (2001) for a discussion of the single nearest neighbor classifier and the cross validation procedure. For this particular feature pair the estimated performance figure of merit is a probability of correct classification of around .76. As will be revealed shortly, this level of performance pales in comparison with that obtained with some of the other feature pairs.

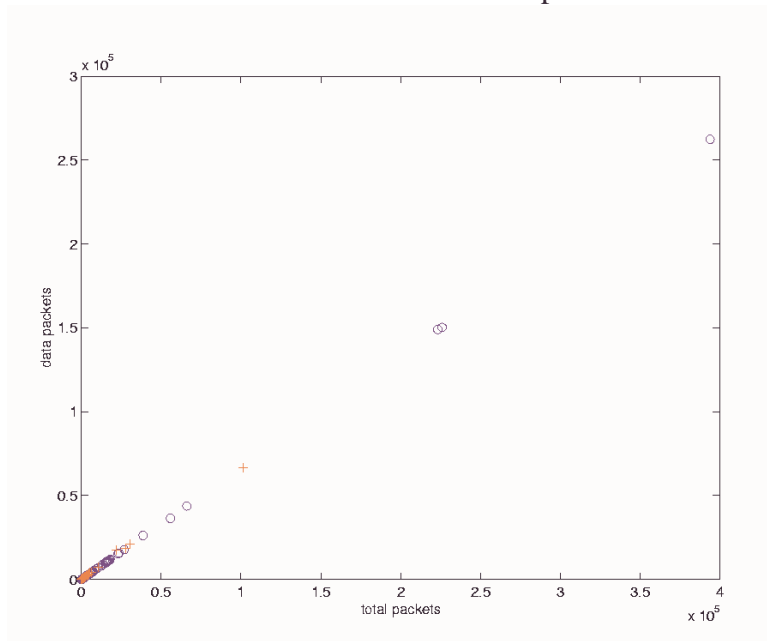


Figure 1: Number of data packets vs. number of total packets where a blue o indicates an scp session and red + indicates a ssh session.

The next feature pair is the absolute value of the difference between the number of data packets out from the server and the number of data packets into the server along with the total number of data packets. The justification of the use of these features is as follows. Consider an scp session wherein a user copies a large file from the server to his client machine. The actual initiation of the scp procedure involves many fewer packets into the server than the number of packets that are transferred from the server. In this sense one would expect the ratio of the absolute value of the difference in the number of data packets out - the number of data packets in to the total number of data packets would be larger in the case of the scp sessions. This expectation is supported by Figure 2. This feature pair seems to offer more hope with regards to distinguishing between the two classes and this hypothesis is supported by the measured single nearest neighbor cross validated performance figure which is .95.

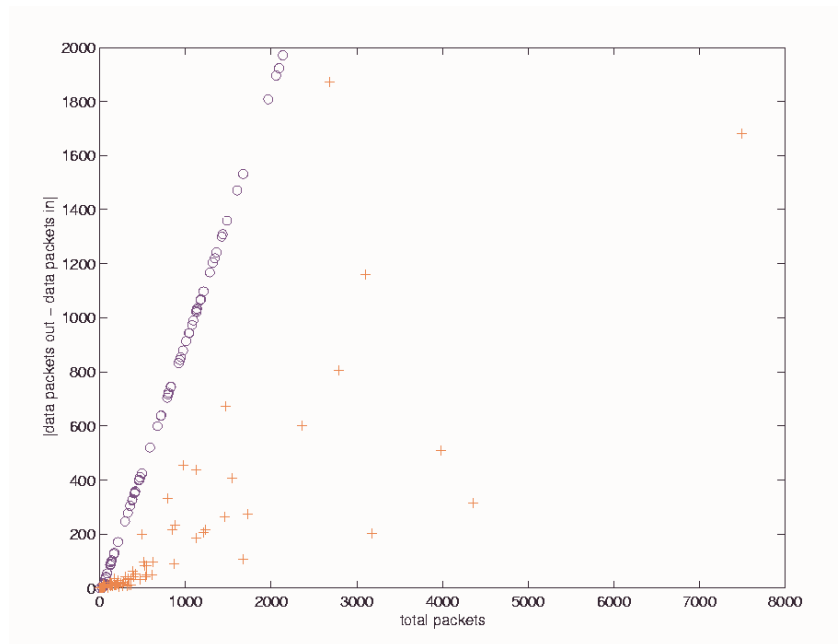


Figure 2: Absolute value of the number of data packets out of the server minus the number of data packets into the server vs. the total number of data packets. The scp observations are plotted as blue circles while the ssh observations are plotted as red plus signs.

The next figure combination that we consider is the natural logarithm of the number of data packets out of from the server in conjunction with the log of the number of data packets into the server. We expect this feature to provide a high discriminatory power because in an ssh session there is more of a one to one exchange of packets between the client and server machines while in a scp session there is a more dichotomous relationship. An initial set of client provided packets leads to a large number of packets solicited from the server. For example if one were using scp to copy a large file from the server then there would be many more packets out of the server than into the server. This speaks to the fact that this feature is directional in nature but this does not in practice present a problem to the analyst in that the direction of the packet flow is know as part of the session information.

Figure 3 presents a plot of the natural log of the number of data packets out from the server vs. the natural log of the number of data packets into the server. The scp sessions are plotted as blue circles in the plot while the ssh sessions appear as red plus signs. One can distinguish a clear separation between the observations associated with these two types of sessions and this perceived separation is supported by the estimated cross validated probability of single nearest neighbor classification performance of .98.

We do note that the discriminatory power decreases in the case where there are not many packets associated with the session. An scp example of this would be when a user copies a very small file. In this case the plot for the scp observations and the ssh observations seems to converge at a v-shaped structure. The close proximity of the scp and ssh observations in this case clearly indicates the lack of discriminatory power and this behavior is in keeping with our intuition.

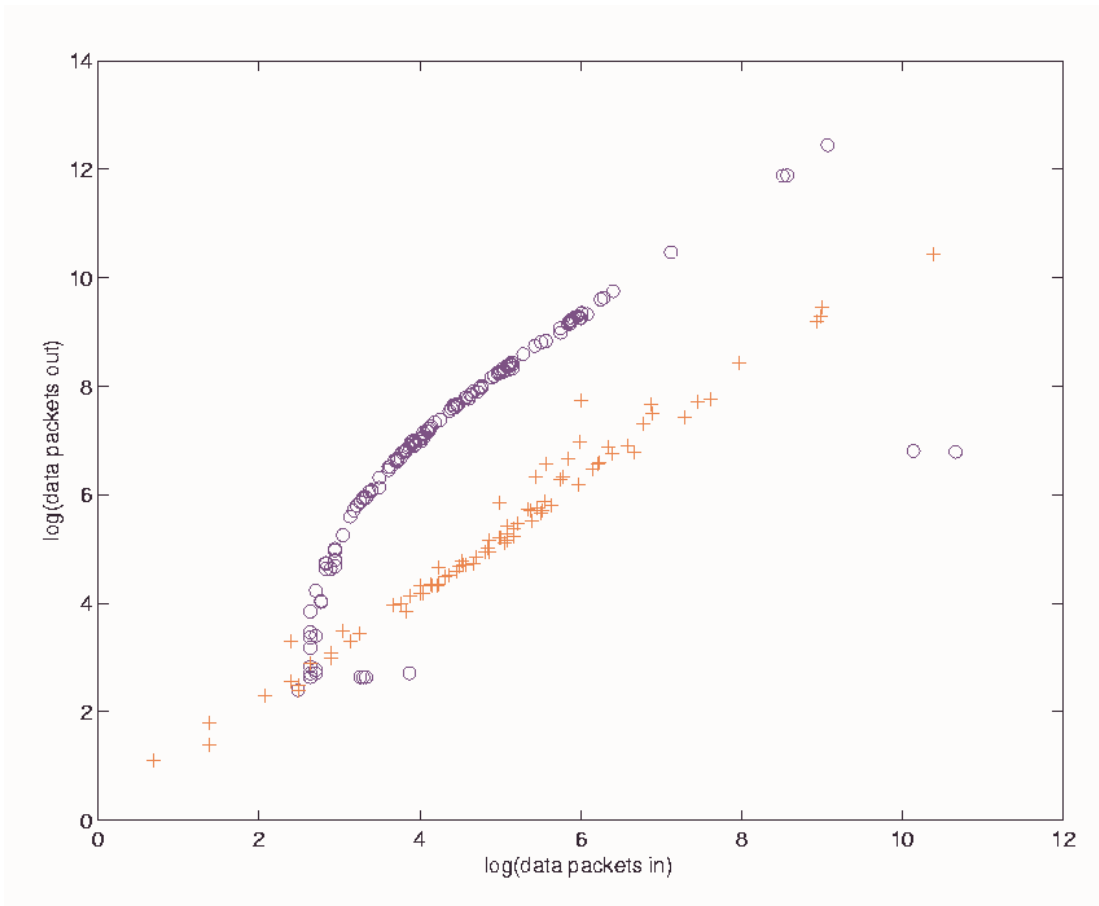


Figure 3: Natural log of the number of data packets out of the server vs. the natural log of the number of data packets into the server. The scp sessions are plotted as blue circles while the ssh sessions are plotted as red plus signs.

The next feature pair that we will examine is the natural log of the number of data bytes along with the natural log of the duration of the session. We would expect given the same duration for the session that the number of data bytes transferred during a scp session would be much higher than

the number of data bytes transferred during a ssh session. A human typist would not be expected to be able to keep up with the associated transfer rates that one would expect to exist during an scp session. This assumes of course that the user is not involved in tunneling any sort of traffic through the ssh pipe. In this case one would expect this measured feature pair to fall somewhere between the pure ssh xterm type session and the scp totally machine controlled session. Figure 4 presents a plot of the natural logarithm of the number of data bytes vs. the natural logarithm of the duration of the session. Once again we have plotted the scp observations as blue circles and have plotted the ssh observations as red plus signs. As expected the scp observations in general lie above the ssh observations as expected. This observation is supported by the measured performance metric of .95.

We point out for completeness that this feature suffers from the same lack of discriminatory power in the case of short duration sessions. It is not surprising in this case to note that the scp observations and the ssh observations are very close in the case of short duration sessions. In this case the number packets transferred as part of an scp session and the number of packets transferred as part of an scp session are very close.

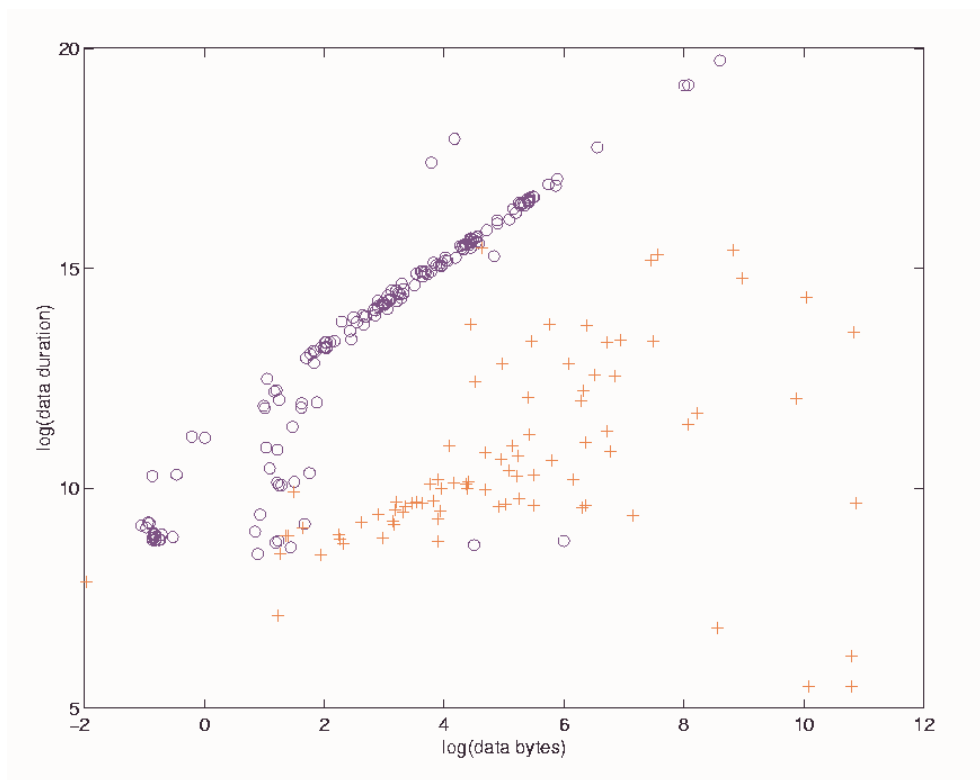


Figure 4: Natural log of the number of data bytes vs. the natural log of the duration of the session. The scp sessions are plotted as blue circles while the ssh sessions are plotted as red plus signs.

4 Conclusions

We have presented a new approach to the discernment of scp and ssh sessions. We have chosen to utilize a number of session-based features for this preliminary exploratory data and discriminant

analysis study. We have examined a number of two-dimensional orthogonal projections of our extracted feature set. We have provided simple scatter plots and one nearest neighbor cross-validated classifier performance for each of the two dimensional feature combinations. These preliminary results suggest that the natural logarithm of the number of data packets out along with the natural logarithm of the number of data packets in is the optimal feature pair among those two dimensional pairs studied. This provided a measured figure of merit of .98.

There are numerous things that were not done as part of this study. First it would make sense to quantify performance using the full set of extracted features at least those features that are appropriate to include. Second it would make sense to examine linear combinations of the features. The astute reader might wonder why we have not provided results related to the Fisher Linear Discriminant procedure for example. It would also make sense to examine the features within a hyperdimensional framework such as parallel coordinates. In this manner we could ascertain the overlap between the observations in the full feature space and possibly rapidly determine those feature combinations that would possess the most discriminatory power. This parallel coordinates framework could be integrated with a procedure to explore linear projections such as the grand tour of Asimov. This would provide a convenient framework to identify fortuitous linear combinations of the features.

There are two other issues that we have not discussed due to the page limitations of this paper. The first is the problem of distinguishing between various users based on their ssh or scp utilization patterns. The second issue is the detection of tunneled applications. One can ssh into a server and then tunnel any X-Windows type of application. Although this is not normally done, one could tunnel a web session or text editor session such as emacs. The detection and classification of these types of tunneled applications is also of interest but will have to be relegated to future investigations.

We finally note that this analysis has been based on a very small data set. The data set suffers from not only lack of cardinality but also from bias in that the distribution of observations across IP address (user) is not uniform. We hope to address these experimental design issues with future data collections.

5 Acknowledgments

The first two authors (JLS and MLA) would like to acknowledge the support of the NSWCCD ILIR Program along with the support of the Missile Defense Agency. The work of the third author (EJW) was completed under the sponsorship of the Air Force Office of Scientific Research under the contract F49620-01-1-0274 and the Defense Advanced Research Projects Agency through cooperative agreement 8105-48267 with Johns Hopkins University. The authors would also like to thank Dr. David Marchette for initially suggesting this effort, for his many insightful comments during the experimental design and analysis phase of this research, and for organizing this session. Finally the authors would like to thank Mr. Don Talsma for providing us access to the XTRACKER AND TRACKER software.

References

Duda, R. O., Hart, P. E. and Stork, D. G. (2001), *Scene Analysis and Pattern Recognition*, New York: John Wiley and Sons.

Solka, J. L., Marchette, D. J., Wallet, B. W. (2000) "Statistical visualization methods in intrusion detection," *Computing Science and Statistics*, 32, 16-24.

Stevens, W. Richard (1994), *The Protocols (TCP/IP Illustrated, Volume 1)*, Reading, MA: Addison-Wesley.