Correlating TCP/IP Interactive Sessions with Correlation Coefficient to Detect Stepping-Stone Intrusion

Guoqing Zhao  
College of Information Engineering  
Beijing Institute of Petro-Chemical Technology  
19 Qingyuan N. Rd, Beijing, 102617 China  
e-mail: zhaoguoqing@bipt.edu.cn

Long Ni  
Department of Electrical and Computer Engineering  
North Carolina A & T State University  
Greensboro, NC 27401 U.S.A.  
e-mail: nilong@ncat.edu

Jianhua Yang, Gurdeep S. Hura  
Department of Math & Computer Science  
University of Maryland Eastern Shore  
Princess Anne, MD 21853 U.S.A.  
e-mail: {jyang1, ghura}@umes.edu

Shou-Hsuan Stephen Huang  
Department of Computer Science  
University of Houston  
Houston, TX 77204 U.S.A.  
Email: shuang@cs.uh.edu

Abstract — Most network intruders launch their attacks through stepping-stones to reduce the risks of being discovered. To uncover such intrusions, one prevalent, challenging, and critical way is to compare an incoming connection with outgoing connections to determine if a computer is used as a stepping-stone. In this paper, we present a way by using signal processing technology — correlation coefficient, such as Spearman Rank, Kendall Tau Rank, and Pearson Product-Moment, to correlate two sessions to identify stepping-stone intrusions. The contribution of this paper is that we are the first one to apply correlation coefficient to stepping-stone intrusion detection, and more importantly, it is not necessary to monitor a session for a long time to conclude a stepping-stone intrusion. The experiment results showed that a stepping-stone intrusion can be detected while an intruder input the username and password. Further work needs to be done to test if this approach could resist intruders' evasion.

Keywords – stepping-stone; intrusion detection; network security; correlation coefficient; interactive session

I. INTRODUCTION

It is not a secret that most intruders usually attack other computers through stepping-stones [1]. One obvious reason is that using stepping-stone could make the intruders safe from being detected, even captured. With the development of computer technologies many approaches to detect stepping-stone intrusion were proposed [1], such as Content-Based Thumbprint [3], Time-Based Approach [1], Deviation-Based Approach [4], Round-Trip Time Approach [5, 2], and Packet Number Difference-Based Approach [6, 7]. Usually intruders take advantage of the vulnerabilities of TCP/IP to manipulate TCP sessions in order to avoid detection. Most commonly used manipulations are time-jittering and chaff-perturbation [6]. Some approaches have been developed to resist intruders’ manipulations, such as estimating the length of a connection [8], and matching TCP/IP packets [9].

The objective of this paper is to refine the developed algorithms to achieve enhances detection performance for stepping-stone intrusion. We developed one method to process network traffic (TCP/IP packets) by computing the packets to determine the occurrence of a stepping-stone intrusion [10]. Both theoretical analysis and simulation studies confirmed the effectiveness and feasibility in detecting stepping-stone intrusion by monitoring and analyzing network traffic.

To make the detection method [10] more stable, practical, and efficient, refinement is extended in this paper. Note that the approach “matching TCP/IP packets” [10] methods simply compared the number of Send and Echo packets to determine if there exists stepping-stone intrusion. The potential problem of this method is that once a session is manipulated, more packets need to be monitored. That means this method is not efficient against intruders’ intrusion. Moreover, this method is also not so practical because lots of packets need to be collected. More importantly, counting packet number or matching Send and Echo packet are not extremely reliable characteristics of the Internet traffic. The chances are still high that traffic with different individual packets happens to have the same total packet number, or high matching rate. Therefore, it is of practical interest to develop innovative approach to avoid problems of matching TCP/IP packets to detect stepping-stone intrusion more effectively.

In this paper, we propose a new approach that applies correlation coefficient theory [11, 12, 13] to stepping-stone intrusion detection. In this approach, we consider collected packets as a signal, and it would help to apply existing signal processing and analysis theory, such as correlation coefficient, to addressing the issue of stepping-stone intrusion detection. In this paper, we only focus on correlating incoming and outgoing connections to detect...
The raw scores are converted to ranks, and the differences \( d \) between the ranks of each observation on the two variables are computed. If there are no tied ranks, i.e., then \( \rho \) is given by:

\[
\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}
\]

where \( d_i \) is the difference between each rank of corresponding values of \( x \) and \( y \), and \( n \) is the number of \( (x, y) \) pairs.

B. Kendall Tau Rank

The Kendall Tau coefficient \( \tau \) [11] has the following properties: if the agreement between two rankings is perfect, i.e., the two rankings are the same, the coefficient has the value 1; if the disagreement between the two rankings is imperfect, i.e., one ranking is the reverse of the other, the coefficient has the value \(-1\); for all other arrangements the value lies in between \(-1\) and 1, and increasing values imply increasing agreement between two rankings. If the rankings are completely independent, the coefficient has the value 0 on average. Kendall Tau coefficient is defined as

\[
\tau = \frac{4P}{n(n-1)} - 1
\]

where \( n \) is the number of items, and \( P \) is the sum of all the items ranked after a given item by both rankings. \( P \) can be also interpreted as the number of concordant pairs. The denominator in the definition of \( \tau \) can be interpreted as the total number of item pairs. So, a high value of \( P \) means that most pairs are concordant, indicating that the two rankings are consistent. Note that a tied pair is not regarded as concordant or discordant. If there are a large number of ties, the total number of pairs (in the denominator of the expression of \( \tau \) ) should be adjusted accordingly.

C. Pearson Product-Moment

The Pearson product-moment [12] is the best known correlation coefficient, which was first introduced by Francis Galton. This correlation coefficient is computed by dividing the covariance of the two variables through the product of their standard deviations. Mathematical properties of Pearson product-moment correlation coefficient are described as following.

The correlation coefficient \( \rho_{x,y} \) between two random variables \( X \) and \( Y \) with expected values \( \mu_x \) and \( \mu_y \), and standard deviations \( \sigma_x \) and \( \sigma_y \) is defined as:
\[
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y}
\]

where \(E\) is the expected value operator, and \(\text{cov}\) represents covariance. Since \(\mu_X = E(X)\) and \(\sigma_X^2 = E(X^2) - E^2(X)\) and likewise for \(Y\), then the following holds,

\[
\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}}
\]

\[= \frac{n \sum x_i y_j - \sum x_i \sum y_j}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_j^2 - (\sum y_j)^2}}
\]

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed one in absolute value.

The correlation is 1 in the case of an increasing linear relationship, \(-1\) in the case of a decreasing linear relationship, and some values in between in all other cases, indicating the degree of linear dependence between variables. The closer the coefficient is to either \(-1\) or 1, the stronger the correlation between the variables.

If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. A correlation between two variables is diluted in the presence of measurement error around the estimates of one or two variables, in which disattenuation provides a more accurate coefficient.

D. The correlation algorithm

If we monitor the incoming and outgoing connections of a computer used as a stepping-stone as shown in Figure 1 and collect the Send packets in each connection, we get two sequences \(S_1 = \{s_{11}, s_{12}, ..., s_{1n+1}\}\) from the incoming connection, and \(S_2 = \{s_{21}, s_{22}, ..., s_{2n+1}\}\) from the outgoing connection. We assume that \(S_1\) and \(S_2\) all have \(n+1\) packets, where \(s_i\) represents the timestamp of \(j^{th}\) packet, \(i = 1\) represents incoming connection, and \(i = 2\) represents outgoing connection.

We convert these two sequences to two discrete signals \(f_1(n)\) and \(f_2(n)\) by computing the intervals between \(s_{ij}\) and \(s_{ij+1}\) as the following,

\[
f_1(n) = \{t_{11}, t_{12}, ..., t_{1n}\}
\]

\[
f_2(n) = \{t_{21}, t_{22}, ..., t_{2n}\}
\]

where \(t_i = s_{ij+1} - s_{ij}\) and \(i = 1, 2, j = 1, 2, ..., n\).

Use the above three approaches to compute the correlation coefficients to determine if the two connections are delayed. If they are delayed, stepping-stone intrusion is identified. Otherwise, they are normal pair. The above idea can be summarized as the following correlation algorithm, Correlation Coefficient Detection Algorithm (CCDA), in which \(0 < \varepsilon < 1\) is a threshold.

CCDA \((S_1, S_2, \varepsilon)\):

1. Compute \(f_1(n)\) and \(f_2(m)\) based on \(S_1\) and \(S_2\);
2. Compute \(\rho, \tau, \rho_{X,Y}\) using formula (1), (2) and (3) respectively;
3. Compute the minimum \(\sigma\) of the differences between one and each of the three coefficients computed in step 2,
   \[
   \sigma = \min \{1 - \rho, 1 - \tau, 1 - \rho_{X,Y} \};
   \]
4. Check if inequality \(\sigma < \varepsilon\) satisfied. If ‘YES’, return Stepping-stone pair; if ‘NO’, return normal pair.

End

III. EMPIRICAL STUDY

A. Stepping-Stone Traffic Monitoring and Recording

In this experiment, a connection chain was established by using OpenSSH starting from a local host ‘\(H_1\)’ in Hampton, VA, USA, then login to another host ‘Acl09’ in National Institute of Aerospace, Hampton, VA, USA, finally from there login to a remote host ‘Acl08’ that is an assumed victim site located in Huston, Texas, USA. The connection chain was \(H_1 \rightarrow Acl09 \rightarrow Acl08\) as shown in Figure 1. The local host, Acl09, was used as the monitoring host, and Acl08 was the remote victim host. All the incoming and outgoing connections at Acl09 were monitored and recorded into files by a TCP/IP packet capturing program written with C++ language from the beginning to the end of the session.

![Figure 1. Traffic monitoring in a host](image)

Similar experiments were repeated many times in many days, and all the Send and Echo packets coming in and going out the monitoring unit, local host Acl09, were monitored, collected, and recorded in a file separately. These data were used for later analysis and computation of correlation coefficient. All the sessions or connections including incoming and outgoing connections in this host were recorded in files simultaneously.
Both the source $H_i$ and the monitoring center Acl09 were located in Hampton, VA. An outgoing connection was established between Acl09 and the destination Acl08 located in Houston, TX. In truth, the outgoing and incoming connections formed a relayed connection pair starting from $H_i$ and ending at Acl08. This made the monitoring sensor Acl09 a stepping-stone host during the whole experiment. In CCDA algorithm, we only use Send packets sequence. They are the send packets from source $H_i$ to stepping-stone Acl09 and the send packets from Acl09 to destination Acl08.

B. Stepping-Stone Detection Experiment

In this experiment, the data transmitted between $H_i$ and Acl09, and the data between Acl09 and Acl08 were recorded. They were divided automatically by a program into four packet streams: two incoming and two outgoing or two Send packets. Meanwhile, the time intervals were computed for each Send packet stream. The time intervals of these two Send packet streams are clearly shown in Figure 2. Note the time stamps are replaced by time intervals. Therefore, these figures show the values of intervals between any adjacent Send packets from the beginning to the termination of the connection. Note packet size was not used for comparison, because the experiments showed most of the packets of a connection have almost the same size, and different connections have different packet sizes. Therefore, packet size is not an ideal parameter to distinguish one packet stream from another. The simulation results suggested that time interval is a perfect characteristic to distinguish one Send packet stream from another.

Figure 2 shows the similarity of the packet interval pattern between the incoming Send packets (from source $H_i$ to Acl09) and the outgoing Send packets (from Acl09 to Acl08). Note that the incoming Send packets are longer than outgoing Send packets, and the outgoing Send packets only started some time later and ended some time ear-

C. Correlation Coefficient Results

We computed the three correlation coefficients for the above traffics in Figure 2 recorded from the experiments. The computing results are shown in Figure 3-6. It is shown clearly in these figures that only the relayed connections have high values for each of the three correlation coefficients. When there was no stepping stone, i.e., before the outgoing connection was established or after it
was terminated, the correlation coefficient value is extremely low. Note that the value of any of the three correlation coefficients is very small for normal pairs (unrelayed connections).

Notice although at some points where there were no relayed connections, one of the three correlation coefficients may still have high value, which is referred to false positive alarm. However, at that point at least one (or both) of the other two coefficients is small enough (usually less than 0.1 in a scale of 1.0 as maximum). Therefore, if these three correlation coefficients are used together, the results could be much better. From the experiments, we did not find any false negative alarms which maybe exist. The reason is all these three coefficients are effective in detecting any similarity between two sequences.

The relayed connection pairs have enough similarity to be detected by the correlation coefficients. However, there are few exceptions where the small similarity between normal pairs leads to a big correlation coefficient value. This is because that these coefficients are not specially designed for detecting the similarities of Send packet sequences. They could be used to detect some complicated similarities which are not the similarities at all in our application. Those exceptions could be solved by computing the smallest one of the three correlation coefficients. The reason is that we found that for most exceptions there is always at least one coefficient which can remain small value. Figure 3-5 shows clearly this pattern. Therefore, as shown in Figure 6, every experiment was repeated three times and we computed the three correlation coefficients respectively and only kept the minimum. The possibility of false positive alarm was largely reduced.

The coefficient value of relayed connection pairs is affected little by this strategy. However, the values of normal connection pairs were lowered dramatically, thus this strategy effectively improves the performance of stepping-stone intrusion detection algorithm CCDA. The benefit of minimum-picking strategy can be demonstrated as following. Suppose the number of total experiments is $n$ for each correlation coefficient, and $k_1, k_2, \text{ and } k_3$ are the numbers of false positive alarms for Spearman, Kendall Tau, and Pearson product-moment correlation coefficients, respectively. Then the false positive alarm rate for each coefficient is as following.

$$R_{fpa} = \frac{k_i}{n}, \quad i = 1, 2, 3$$

If the correlation coefficient is repeated three times with Spearman, Kendall Tau, and Pearson product-moment correlation coefficient, respectively, the total false positive alarm rate will be reduced to as the following,

$$R_{fpa} = \left( \frac{k_1}{n} \right) \left( \frac{k_2}{n} \right) \left( \frac{k_3}{n} \right)$$

Previous simulation results show $R_{fpa} = 1/15$ when $n = 15, k_1 = k_2 = k_3 = 1$ while one coefficient is used. If three coefficients are used, $R_{fpa}$ would be reduced to $1/15^3$. Therefore, the performance is largely improved by computing all the three coefficients and picking up the minimum.

IV. RELATED WORK

There have been a number of stepping-stone detection methods using correlation analysis recently. Most of them were just using the idea of correlation to find some kind of similarities between traffic of different connections. Actually, they are not related to signal processing based correlation or statistic based correlation coefficient. A recent approach is called multi-dimensional flow correlation [14, 15], which uses multiple characteristics of a packet flow to
conduct the correlation analysis instead of just one parameter. These characteristics include packet event times, packet inter-arrival times, byte per packet, cumulative bytes per packet, bytes per burst, and periodic throughput samples. It is a good idea to use multiple characteristics. However, this approach does not prove the advantages of using multiple characteristics.

There are very few approaches in literature which focus on the seemingly best characteristic for correlation: time interval of adjacent packets. Multi-scale Stepping-Stone Detection [6] is the one which uses wavelet coefficients to compute the correlation coefficient based on the time intervals of adjacent packets. The problem is that this method only focuses on the theoretical analysis, it is not implemented. Moreover, it needs high computational cost.

Inter-Packet Delay Based Correlation approach [16] is the only one available in literature which addresses the correlation coefficient and focuses on the time intervals of adjacent packets. It proposed four correlation point functions: 1) min/max sum ratio; 2) statistical correlation; 3) normalized dot product 1; 4) normalized dot product 2. It pointed out through experiments that min/max sum ratio has the best performance, and the statistical correlation is the worst. The promising aspect of this method is it could obtain correlation coefficient using only a few dozens of packets, and achieve relatively high performance with only a few dozen of packets. It showed that the time intervals of adjacent packets is a good choice, which makes it possible to distinguish different packet traffics based on very few collected packets, and meantime maintain a high accuracy of the detection.

V. CONCLUSION AND FUTURE WORK

In this paper we have proposed a way using signal processing approach—correlation coefficient to detect stepping-stone intrusion. The experiment results showed that it is more efficient than others because it does not have to monitor connections a long time. We also found that the way to reduce false positive error by picking up the minimum of the three correlation coefficients. Theoretically this error can be reduced cubically from $1/n$ to $1/n^3$ where $n$ is the experimental times.

The future direction will take into account the following two facts regarding the inconvenience for intruders’ evasion. One is that the upper bound of the holding time for time-jittering evasion exists due to the fact that the users (hackers) of the interactive connection chain are human, and it is impractical or impossible for normal people to have to wait a long time before receiving the feedback of their requests. Theoretically, it is a reasonable assumption that there exists a maximum tolerable delay for any hacker to impose [6]. Any period longer than this limitation is impractical for any normal person to implement. This makes the extent of complication of time-jittering bounded and is encouraging for the success of the future approaches intended to resist intruder’s time-jittering evasion.

REFERENCES