Statistical Traffic Modeling for Network Intrusion Detection

João B. D. Cabrera, B. Ravichandran and Raman K. Mehra
Scientific Systems Company
500 West Cummings Park, Suite 3000
Woburn, MA 01801
cabrera,ravi,rkm@ssci.com

Abstract

This paper examines the application of Statistical Traffic Modeling for detecting novel attacks against computer networks. We discuss the application of Network Activity Models and Application Models using the 1998 DARPA Intrusion Detection Evaluation dataset. Network Activity Models monitor the volume of traffic in the network, while Application Models describe the operation of application protocols. By plotting the ROC (Receiver Operating Characteristic) curves induced by the traffic activity, we quantify the effectiveness of Network Activity Models in discriminating normal connections from attack connections generated by Denial-of-Service and Probing attacks. It is verified that Denial-of-Service and Probing attacks leave traces on simple Network Activity Models, with rates of false alarm which are comparable to the false alarm rates obtained by the participants of the 1998 DARPA Evaluation, in which much more complex detection schemes were utilized. For Application Models, we use the Kolmogorov-Smirnov Test to show that attacks using telnet connections in the DARPA dataset form a population which is statistically different from the normal telnet connections. The statistics used in our study are the number of bytes from the responder, the byte ratio responder-originator. Again, our results are comparable to those obtained in the DARPA Evaluation.

1. Introduction

Intrusion Detection Systems (IDSs) extract information from a computer or a network of computers, and attempt to detect the presence of intrusions from external sources, as well as system abuses by authorized users. The detection of intrusions or system abuses (both called attacks) presupposes the existence of a model. Modeling can take two forms. In misuse detection, one models the known attack patterns through the construction of a library of attack signatures. Incoming patterns that match an element of the library are labeled as attacks. If only exact matching is allowed, misuse detectors operate with no false alarms. By allowing some tolerance in attack matching, one runs the risk of false alarms, but expects to be able to detect certain classes of unknown attacks, that “do not deviate much” from the attacks listed in the library. We call such unknown attacks neighboring attacks. In anomaly detection, one models the normal behavior of the system. Incoming patterns that deviate substantially from normal behavior are labeled as attacks. Implicit in the utilization of abnormal patterns to indicate attacks is the premise that malicious activity is a subset of anomalous activity. One is tacitly accepting the presence of false alarms in this case, in exchange for the hope of detecting unknown attacks, which may be substantially different from the neighboring attacks; we call these novel attacks.

Detecting novel attacks while keeping acceptably low rates of false alarm is possibly the most challenging and important problem in Intrusion Detection ([14]). We argue in this paper that features associated with traffic modeling (volume of traffic in the network, and statistics for the operation of application protocols) are particularly suited for detecting general novel attacks. We examine the role of Statistical Traffic Modeling (eg. [5], [12], [13]) in the detection of novel attacks. We will not be designing classifiers as described in [4] and [8], but evaluating the discriminating power of traffic measurements in a sense detailed in sections 2.2, 4 and 5. We describe the extent by which Network Activity Models and Application Models, capturing the volume of traffic and the operation of individual application protocols can be used to discriminate between normal and under-attack states on a network. Network Activity Models are used to detect the presence of Denial-of-Service and Probing Attacks, while Application Models are used to detect the presence of attacks which utilize a given application. We study the performance of these models using the datasets from the 1998 DARPA Intrusion Detection Evaluation. By plotting the ROC (Receiver Operating Characteristic) curves induced by the traffic activity, we show that Denial-of-Service and Probing attacks leave traces on simple Network Activity Models, with rates of false alarm which are comparable to the false alarm rates obtained by the participants of the 1998 DARPA evaluation. Traffic monitoring is performed at the time scale of connection arrivals. We also show that attacks using telnet connections in the
DARPA dataset form a population which is statistically different from the normal telnet connections observed in the DARPA dataset. The Kolmogorov-Smirnov Test is used for comparing the normal and attack populations.

2. Statistical Background

2.1. Poisson processes

Let \( X(t) \), \( 0 \leq t < \infty \), denote the number of events occurring in the time interval \([0, t]\). For \( 0 \leq s \leq t \) the random variable \( X(t) - X(s) \) denotes the number of events occurring in the time interval \([s, t]\). If the waiting time between successive events are independent and exponentially distributed with common parameter \( \lambda \) then \( X(t) \) is a Poisson process. The following properties are important in our discussions in section 4 (eg. \([1]\)):

P1 \( X(t) - X(s) \) has a Poisson distribution with parameter \( \lambda(t - s) = \lambda t \) for all \( 0 \leq s \leq t \leq \infty \), i.e. \( P[X(t) - X(s) = k] = \frac{e^{-\lambda t} (\lambda t)^k}{k!} \).

P2 \( X(t_2) - X(t_1), X(t_3) - X(t_2), \ldots, X(t_n) - X(t_{n-1}) \) are independent for \( 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \).

P3 If \( X_1(t), X_2(t), \ldots, X_n(t) \) are independent Poisson processes with parameters \( \lambda_1, \lambda_2, \ldots, \lambda_n \) then \( Y(t) := \sum_{i=1}^{n} X_i(t) \) is a Poisson process with parameter \( \lambda = \sum_{i=1}^{n} \lambda_i \).

2.2. ROC curves and discrimination power

Suppose the input \( \alpha \) to a detector assumes values in the range \( 0 \leq \alpha \leq \alpha_{\text{max}} \). The ROC (Receiver Operating Characteristic) \([15]\) curve induced by \( \alpha \) is obtained by computing the rate of false alarms and detections for each input value in the range. ROC curves are used in this paper in two situations. In section 4 the variable of interest is the total number of connection arrivals on a given interval. There is a perfect analogy with the radar detection paradigm in this case. The noise in this case corresponds to the normal network traffic, while the signal corresponds to the presence of an attack. In section 5 the variables of interest are the number of bytes from the responder \( X_r \), and the byte ratio responder-originator \( \rho \) observed in telnet connections. One finds that both \( X_r \) and \( \rho \) are much lower on average for attack populations. The signal-noise radar analogy is not valid in this case, but we can still obtain the ROC curves induced by these two variables in the normal and attack populations.

2.3. The Kolmogorov-Smirnov Test

Our exposition in this section closely follows the developments in \([11]\). In using the Kolmogorov-Smirnov Test, a comparison is made between a theoretical, completely specified target cdf \( F_T(z) \) and a sample cdf, \( F_n(z) \). The sample (or empirical) cdf is defined as follows:

**Definition 1 (Empirical Distribution Functions)**

Let the jump function \( t(u) \) be defined as \( t(u) = 0 \) if \( u < 0 \), and \( t(u) = 1 \) if \( u \geq 0 \). If \( z_1, z_2, \ldots, z_n \) constitute a random sample from a population with cdf \( F(z) \), the sample (or empirical) cdf \( F_n(z) \) is defined as \( F_n(z) = \frac{1}{n} \sum_{k=1}^{n} t(z - z_k) \).

**Definition 2 (Kolmogorov statistic)**

Given a sample \( z_1, z_2, \ldots, z_n \) drawn from a cdf \( F(z) \), the Kolmogorov statistic of the sample \( D_n(z_1, z_2, \ldots, z_n) \) is defined as: \( D_n := \sup_{z \in Z} |F_n(z) - F(z)| \), where \( Z \) denotes the sample space for \( z \).

The usefulness of \( D_n \) for statistical inference stems from the fact that if \( F(z) \) is continuous, then the distribution of \( D_n \) does not depend on \( F(z) \), only on the sample size \( n \). This is a very important property for the class of problems investigated here, for which very little is known about the underlying processes generating the data, especially in the case of attacks. To describe the Kolmogorov-Smirnov test it is necessary to have a notation for the quantiles (or percentage points) of the distribution of \( D_n \). Let \( D_{\alpha} \), satisfy \( P(D_n \leq D_{\alpha}) = \alpha \) when \( D_n \) is based on a random sample drawn from a continuous population. Recalling the definition of \( D_n \), we have with probability \( \gamma \)

\[-D_n \leq F_n(z) - F(z) \leq D_n, \quad \text{for all } z \in Z. \quad (1)\]

Equation (1) can be rewritten in two different ways:

\[ F(z) - D_n \leq F_n(z) \leq F(z) + D_n, \quad (2)\]

\[ F_n(z) - D_n \leq F(z) \leq F_n(z) + D_n, \quad (3)\]

Equation (2) is the motivation for the Kolmogorov-Smirnov test. It says that the probability that, under random sampling from a population with continuous distribution \( F(z) \), the sample cdf \( F_n(z) \) deviates from \( F(z) \) anywhere (i.e. for any \( z \in Z \)) by more than \( D_n \) is given by \( \alpha = 1 - \gamma \). This observation leads to the test:

**Definition 3 (The Kolmogorov-Smirnov Test)**

Given a completely specified distribution function \( F_0(z) \), and a sample \( z_1, z_2, \ldots, z_n \), the Null Hypothesis for the test is: \( H_0: F(z) = F_0(z) \), for all \( z \in Z \). The alternative hypothesis is clearly: \( H_1: F(z) \neq F_0(z) \) for some \( z \in Z \). We reject the \( H_0 \) at significance level \( \alpha \) if \( D_n = \sup_{z \in Z} |F_n(z) - F(z)| > D_{1-\alpha} \).

Tables for values of \( D_{\alpha} \) are available in \([3]\), p. 430 for \( \alpha = 1\%, 2\%, 5\%, 10\%, \) and \( 20\% \). At first sight the alternative hypothesis seems very weak, since it is virtually impossible that raw data could ever match a theoretical distribution exactly. To better interpret the alternative hypothesis, we resort to equation (3). Define at first

\[ L_n(z) = \max \{0, F_n(z) - D_n\}\]

\[ U_n(z) = \min \{1, F_n(z) + D_n\}\]
Equation (3) implies that under random sampling from a population with continuous distribution $F(z)$ the probability is $\gamma$ that the piecewise constant functions $L_n(z)$ and $U_n(z)$ contain $F(z)$ between themselves for all $z \in Z$. The functions $L_n(z)$ and $U_n(z)$ are called lower and upper confidence contours for $F(z)$. The region between $L_n(z)$ and $U_n(z)$ is called the confidence band for $F(z)$. The probability $\gamma$ connected with the sampling process underlying equation (3) is called the confidence coefficient associated with this confidence band. If $H_0$ is rejected with significance level $\alpha$, it means that all the other $H_0$ for which the corresponding $F_0(z)$ satisfy $L_n(z) \leq F_0(z) \leq U_n(z)$ are also rejected with confidence level $\alpha$. Hence, if we reject $H_0$ we reject the whole family $H_0$, and accept the whole family of alternative hypothesis $H_1$.

3. The 1998 DARPA Evaluation

The dataset used in this study was collected for the DARPA 1998 Offline Intrusion Detection Evaluation ([6], [7], [10]). The complete dataset for the 1998 evaluation consists of 7 weeks of training data and 2 weeks of testing data on an experimental testing network. For each of the connections, the following information is supplied: date; start time; duration; service; source and destination ports; source and destination IP addresses; attack/non-attack label. We call these the intrinsic variables for the connection. In the case of the attacks, the type of the attack is also supplied. In our study, we characterize the discriminative capabilities of certain traffic variables during weeks 3 to 7, corresponding to the training data. There are four classes of attacks in the dataset, with various types on each class:

- **DOS** (Denial-of-Service) attacks are effected by loading a computing or memory resource with a large number of requests, thus rendering the resource too busy or too full to handle legitimate requests. By their very nature, DOS attacks cause an increase on network activity, since the attacker needs to establish a large number of connections in order to attain its objective. **PROBE** attacks make use of a growing number of utilities which automatically scan a network of computers to find vulnerabilities. As described in section 4, DOS and PROBE attacks generate network activity which can be potentially detected by an Intrusion Detection System. **U2R** (User-to-Root) U2R attacks are those in which an attacker with legitimate access to a user account on the system (gained by sniffing passwords, social engineering, etc.) uses this access to exploit some vulnerability to gain root access to the system. U2R attacks can be run from any interactive user session, by sitting at the console or remotely, through telnet or rlogin. If the attacker is sitting at the console, the only way to catch the attack is through user profiling, or by pattern matching to known attacks. If the attacker is operating remotely, the attack session can be recorded by a sniffer. For the DARPA dataset, we observed that the majority of U2R attack types in weeks 3-7 (five out of seven) was registered in telnet connection. Finally, **R2L** (Remote-to-Local) attacks occur when an attacker who has the ability to send packets to a machine over a network exploits some vulnerability to gain local access as a user of the machine. Most R2L attacks explore vulnerabilities on the authentication mechanisms of certain applications, such as ftp, telnet, http, smtp, etc.

There are also **Scenario** attacks, combining elements of the four classes into an attack with multiple phases. Finally, the DARPA dataset also contains **Anomalies** which are certain users behaving abnormally, but not in a malicious way.

4. Network Activity Models

4.1 Motivation

Network Activity Models are motivated by the observation that DOS and PROBE attacks cause an increase in the total number of connections initiated during a given interval. We introduce the following notation:

**Definition 4 (Counting connection arrivals)**

\[ X^T_n (k) : \text{Total number of normal connections which were initiated in the interval } (kT, (k+1)T). \]

\[ X^A_n (k) : \text{Total number of attack connections which were initiated in the interval } (kT, (k+1)T). \]

\[ X^C_n (k) : \text{Total number of connections which were initiated in the interval } (kT, (k+1)T). \]

Clearly \[ X^C_n (k) = X^T_n (k) + X^A_n (k). \]

The time line runs from 8 am to 6 am of the next day, or the last recorded connection, in the days when the network was stopped earlier. Figure 1 depicts the evolution of $X^T_n (k)$ and $X^C_n (k)$ for four choices of sampling interval. Similar behavior can be observed for other days. The following observations can be made:

1. At $t=12:46:20$ there is a DOS attack - **teardrop** - which can be clearly observed in the time scales of 10 seconds and 100 seconds. For the time scales of 1000 seconds and 3600 seconds it could be mistaken by a natural increase of network activity. The **teardrop** attack consists of sending a large number of misfragmented ICMP packets to a host. In this case, over 200 packets (or ICMP "connections") were sent on an interval of just 2 seconds.

2. At $t=22:42:37$ there is a U2R attack - **eject** - which is not noticeable at any time scale. The **eject** attack is effected through a single telnet connection; it is evident that different techniques are needed to detect such attacks, as described in section 5.
3. At t=23:46:35 there is a PROBE attack - *ipsweep* - which is evident at all time scales, but specially at 100, 1000 and 3600 seconds. The *ipsweep* attack is effected by a single source sending many ICMP ping packets to several machines in the network. *ipsweep* related packets are sent over a long interval; in this case the last packet is sent at t=1:16:19, so we have an attack lasting for one and half hours.

4.2 Traffic modeling

It has been reported in [12] and [13] that connections arrivals for user initiated services such as telnet and ftp can be well-modeled by Poisson processes with hourly rates. Based on these and other studies, diurnal patterns of activity and Poisson session arrivals were identified in [5] as good candidates for "Internet invariants". Such invariants are useful to keep in mind while building models that are expected to work under varying circumstances. We have observed the following features in the DARPA dataset:

1. **Diurnal patterns for the normal connections**
   There are roughly three operating regimes for $X_A^{\mu}(k)$. A *day* regime from 8 am to 4 pm, an *evening* regime, from 4 pm to 8 pm, and a *night* regime, from 8 pm to 6 am in the next day. On a first approximation, the *day* and *night* regimes are second-order stationary, with much higher averages for the *day*, as evident from Figure 1. The *evening* regime has a downward trend.

2. **Poisson models for the normal connections**
   For the *day* and *night* regimes, the processes $X_A^{10}(k)$ and $X_A^{100}(k)$ display, on a first approximation, property P2 (section 2.1) associated with Poisson processes.

3. **Heterogeneous behavior for attack connections**
   The DOS and PROBE attacks appear in bursts, with widely varying amplitudes, which are very hard to characterize statistically. U2R and R2L appear single, isolated connections.

Due to space limitations, we refer the reader to our report [2] where these observations are verified in detail. Observation 2. suggests that the processes $X_A^{\mu}(k)$ during the *day* and *night* regimes can be modeled as i.i.d. random variables with pdfs $f_d(x)$ and $f_n(x)$ respectively. Under this simple model, the presence of attacks can be inferred by setting a threshold in the number of observed connections in the interval.

**Remark 1** Notice that we are observing Poisson-like behavior (property P2) for the total number of normal connection arrivals. A possible explanation for this comes from property P3 in section 2.1. The total number of connections is the superposition of all connections across the network (which are individual Poisson processes). Assuming that these individual processes are independent, their sum should also be a Poisson process (P3).

4.3 Attack Intervals and Normal Intervals

In order to evaluate the discriminating power of the connection counting process, we need to define what we mean by normal intervals and attack intervals:

**Definition 5** (Attack intervals and normal intervals)

\[ X_A^\mu(k) > X_A^\nu(k) \quad \Rightarrow \quad \text{[}kT, (k + 1)T\text{]} \text{ attack int.} \]

\[ X_A^\mu(k) \leq X_A^\nu(k) \quad \Rightarrow \quad \text{[}kT, (k + 1)T\text{]} \text{ normal int.} \]

**Remark 2** Once an interval is flagged as an attack interval, there is still the problem of identifying which of the over 50% of the connections are attacks. Another concern is the fact that intervals which are flagged as normal may still have a large number of attack connections. We observed however that intervals containing a large number of connections tend to have a percentage of attack connections very close to 100%. It means that once an interval is flagged as an attack using Definition 5, and the number of connections is moderately large, almost all connections in the interval are attacks. It also means that intervals with a low percentage of attack connections have a small number of connections. We return to this issue in Remark 3. More details are given in [2].

Notice that $X_A^\mu(k)$ and $X_A^\nu(k)$ are not available to the Intrusion Detector. Only $X_A^\mu(k)$ is available. However, since the DARPA dataset labels the normal and attack connections, we can evaluate the discriminating power of $X_A^\mu(k)$ by obtaining the ROC curve induced by $X_A^\nu(k)$ into the normal and attack populations. Conceptually, as described in section 2.2, we will be investigating the performance of the threshold classifier based on $X_A^\nu(k)$ (section 2.2) for all values of the threshold $\alpha$.

4.4 Discriminating power of $X_A^\nu(k)$

To quantify the discriminating power of $X_A^\nu(k)$, we follow the procedure delineated in section 2.2. We obtain different ROC curves corresponding to the *day* and *night* regimes, i.e. we obtain normal and attack populations for the two regimes separately - normal and attack populations are obtained for the 25 days of weeks 3-7. It is evident that a thresholding scheme is not suitable for the *evening* regime. The design of detection techniques for the *evening* regime is not addressed in this paper. The results are presented in Figure 2 where each sub-figure corresponds to a particular regime, and the results for the three time scales are superimposed.

- **day** regime: The results are very good in this case. It is possible to obtain PDs in the range of 60% ($T = 100$ s) to 85% ($T = 1000$ s) with no false alarms.

- **night** regime: The results are substantially better for $T = 10$ s, but inferior overall to those obtained in the *day* regime.
Remark 3 (Relevance for the DARPA evaluation) For the purpose of rating the performance of the algorithms, attacks in [6] are considered as a whole, i.e. each attack counts as one detection, irrespective of the number of sessions. For this reason, DOS and PROBE attacks (several connections) are rated differently than R2L and U2R attacks (one or few attacks). For DOS and PROBE one counts the total number of connections in which an attack was listed, and divide by the total number of connections in the attack. For U2R and R2L an attack is considered to be detected if it was indicated in at least one of the connections. Several DOS and PROBE attacks appear in ten of thousands of connections. The number of normal connections on a typical day is also of tens/hundreds of thousands. In this case, missed detections and false alarms will be scored against a universe of tens of thousands of detections, which explain the extremely low levels of false alarms (less than .2%) reported by the participating teams. Our ROC curves represent intervals, and not individual connections. Hence, we detect an interval as being an attack interval and flag all its connections as attacks we will obtain even better levels of detection \times false alarm in both regimes than those reported in [6]. These remarks are also true for the night regime. Details are given in [2].

5. Application Models

5.1 The telnet connections

Statistics for the telnet connections were obtained from the tcpdump data using the publicly available script tcp-conn. Besides the intrinsic variables for each connection (defined in section 3), we also extracted $X_o$ (no. of bytes from the originator), $X_r$ (no. of bytes from the responder) and the status flag for the connections. The extraction process failed for 6 of the 25 days in the dataset, due to very large file sizes (larger than 4 Gigabytes in some cases). There is a total of 2211 telnet connections in the 19 days. As pointed out in [9], the Termination Status Flag (TSF) is the most important characterization of a connection, since its value is a summary of how the connection has behaved according to network protocols. TSF values besides SF (i.e. normal connection establishment and termination) are of great interest for intrusion detection. The tcp-conn script could not extract $d$ (duration of the connection), $X_o$ and $X_r$ for 776 of the connections; only TSF was available in these cases. Failure of tcp-conn in extracting information from a connection indicates that the connection is corrupted in some way. Table 1 presents a break-down of the telnet connections among these two sets of connections: Set 1 corresponds to the 1435 connections for which $d$, $X_o$, and $X_r$ are available, as well as TSF; Set 2 corresponds to the case where only TSF information is available. It is clear from Table 1 that failure of the tcp-conn in extracting the

<table>
<thead>
<tr>
<th>Set</th>
<th>Total</th>
<th>Normal</th>
<th>Attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 and 2</td>
<td>2211</td>
<td>1468</td>
<td>743</td>
</tr>
<tr>
<td>1: d, $X_o$, $X_r$, TSF</td>
<td>1435</td>
<td>1398</td>
<td>37</td>
</tr>
<tr>
<td>2: TSF only</td>
<td>776</td>
<td>70</td>
<td>706</td>
</tr>
</tbody>
</table>

Table 1. Summary of telnet connections present in weeks 3-7 of the DARPA dataset.

variables $d$, $X_o$, and $X_r$ is strongly related to the presence of attacks. We now investigate the effectiveness of TSF to discriminate between normal and attack telnet connections. We define the Flag Classifier for telnet connections as follows:

Definition 6 (The Flag Classifier)

- The termination flag is SF (Normal Termination) ⇒ The telnet connection is normal.
- The termination flag is not SF (Abnormal Termination) ⇒ The telnet connection contains an attack.

We counted 113 normal connections having abnormal termination (TSF≠SF), and 37 attack connections having normal termination (TSF=SF). Hence, FA and MD rates for the Flag Classifier are:

\[ FA = \frac{113}{1468} = 7.68\%, \quad MD = \frac{37}{743} = 4.98\%, \]

and in fact the termination flag is indeed a very good feature for discriminating normal connections from attack connections. We also notice that all 776 connections in Set 2 (the ones for which tcp-conn failed to extract $d$, $X_o$, $X_r$) had abnormal termination, showing the strong correlation between the presence of attacks, the corruption of the corresponding tcpdump files, and abnormal termination. Notice however that the 37 attack connections in Set 1 had normal termination, i.e. if we apply the Flag Classifier to Set 1 alone, the FA and MD rates become:

\[ FA = \frac{113}{1398} = 8.08\%, \quad MD = \frac{37}{37} = 100\%, \]

which is clearly unacceptable. The crucial question at this stage is: Could these attacks be detected by other means? Or, are distributions of the variables $d$, $X_o$, $X_r$, and $p$ for the attack connections different from the normal connections? Table 2 itemizes the classes of attacks appearing on Set 1 and Set 2. In Set 2, 700 out of the 706 attacks

<table>
<thead>
<tr>
<th>Set</th>
<th>DOS</th>
<th>PROBE</th>
<th>R2L</th>
<th>U2R</th>
<th>S/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>706</td>
<td>701</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Attacks appearing on telnet correspond to the SYN-flow attack neptune. The neptune attack works by creating a large number of partially
opened TCP connections to a machine. This phenomenon is captured by the termination flag of the connection. The fact that the connection is telnet, http, or smtp is immaterial in this case. The 14 U2R attacks appearing in Set 1 correspond to loadmodule (four times), eject (four times), ffb (3 times), perlmagic (two times) format (one time). Hence, five of the seven types of U2R attacks appeared in telnet sessions. Recall from section 3 that the 8 anomaly connections in Set 1 correspond to variations in the usual behavior of particular users, and not to attacks. Hence, we delete these records from the attack population in Set 1, leaving the Attack population with 29 elements. Table 3 gives the mean values for the normal and attack populations in Set 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>No.</th>
<th>$d$</th>
<th>$X_o$</th>
<th>$X_r$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1368</td>
<td>132</td>
<td>409</td>
<td>39220</td>
<td>96</td>
</tr>
<tr>
<td>Attacks</td>
<td>29</td>
<td>59</td>
<td>329</td>
<td>1532</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Table 3. Mean values for telnet connections. As suggested in [12], we compute the geometric mean of the populations.

5.2. Statistical comparison of the normal and attack populations

The hypothesis test in this case attempts to answer the following question: Are the datasets corresponding to attack connections sufficiently different from the datasets corresponding to normal connections? If the answer to this question is yes, it means that the corresponding random variables are useful for discriminating between attacks and normal connections. Therefore, classifiers designed using these variables should lead to good results. The target cdf in this case is the empirical cdf for normal connections. The sample cdf corresponds to the empirical cdf for attack connections. The values of the Kolmogorov statistic for these variables are given in Table 4. Table 5 gives the critical values for the test, for various significance levels, for $n = 29$ (size of the attack population). We can then conclude with a significance better than 1% that for the $X_r$ and $\rho$ variables the attack population does not come from the same distribution as the normal population. Tables 4 and 5 also show an evidence of 10% to 20% for the null hypothesis for $d$, and an evidence of 5% to 10% for the null hypothesis for $X_o$. It means that if we reject the null hypothesis for these variables, and assume that the normal and attack connections are generated by two different processes, then we should be ready to accept probabilities of false alarms of 5%-10% for $X_r$ and 10%-20% for $d$. The Kolmogorov-Smirnov test is conclusive in these two cases. Our general conclusions based on the tests in this section are:

1. There is no conclusive evidence that unscaled $d$ and $X_o$ variables from normal and attack telnet connections correspond to different underlying processes. Hence, $d$ and $X_o$ are not good quantities to discriminate attack connections from normal connections.

2. Main Conclusion: There is very strong evidence that unscaled $X_r$ and $\rho$ variables from normal and attack telnet connections correspond to different underlying processes. This suggests that $X_r$ and $\rho$ are good quantities for discriminating attack connections from normal connections. This suggestion is verified in section 5.3.

5.3 Using $X_r$ and $\rho$ for discrimination: Comparison with DARPA Evaluation

To evaluate the effectiveness of $X_r$ and $\rho$ for discriminating normal and attack telnet connections, we obtain the ROC curves induced by these variables in the normal and attack telnet populations. Figure 3 depicts the results. In order to compare our results with the DARPA 1998 results, we first need to "convert" the FA rates. The testing set used in the DARPA evaluation corresponds to 10 days, with a total of 660,049 normal connections ([6]). According to Table 1 there are 1468 normal telnet connections in 19 days, giving a rate of 77.3 normal telnet connections/day. Hence, in 10 days of the testing we estimate that 773 normal telnet connections are present. Therefore, the FA rates reported in [6] must be multiplied by $\frac{773}{660,049} = 854$ for a fair comparison with our rates. In [6], detection rates are presented for FA rates of 1/day and 10/day. For a fair comparison, a FA rate of 1/day corresponds to 1 false alarm/773 connections, or 1.3%, and FA rate of 10/day corresponds to 13%. The detection rates for the best system among the participants of the DARPA evaluation are given in Table 6. Comparing the results in table 6 (DARPA Evaluation) with the results in Figure 3 show that $X_r$ and $\rho$ alone

<table>
<thead>
<tr>
<th>FA rate</th>
<th>FA rate for telnet</th>
<th>Detection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/day</td>
<td>1.3%</td>
<td>58%</td>
</tr>
<tr>
<td>10/day</td>
<td>13%</td>
<td>72%</td>
</tr>
</tbody>
</table>

gives about the same discrimination capability as the best results in the DARPA Evaluation for a FA rate of 10/day. We find this result extremely encouraging, considering the robustness of these features, and the simplicity of our scheme compared to the techniques used in [8] and [4]. For the FA rate of 1/day, the statistic ρ gives a discrimination capability of about 20% against 58% for the DARPA Evaluation. It suggests that other features are needed if one wants to achieve a higher degree of accuracy.

6. Conclusions

In this paper we examined the performance of simple statistical models for detecting attacks in computer networks. Network Activity Models are motivated by the fact that Denial-of-Service and Probing attacks cause an increase in the total number of connections initiated during a given interval. It is not a surprise for us that these models are effective in flagging the presence of such attacks. The contribution of this study is the quantification of the discriminating capabilities of these models in an isolated fashion. We verified that the rates of false alarm and detection fall in the same range as the ones reported in the DARPA evaluation. We stress that our detectors rely solely on traffic counting statistics across the whole network. This is in contrast with the schemes described on [4] and [8], that utilize host- and service-specific traffic statistics. To illustrate the applicability of Application Models, we have shown that traffic statistics of telnet connections serve as useful discriminants between attacks and non-attacks. We stress that TSF is normal for the whole attack population in this case. In particular, for FA rates of 10/day, our results are of the same order of the results obtained using much more sophisticated methods ([8] and [4]). Although further validation is needed using larger datasets, our results strongly indicate the importance of Statistical Traffic Models as a first line of detection for different classes of intrusions. Important research themes are the evaluation of Application Models for the http service, the combination of information at different scales to improve the accuracy of detectors, and the formulation of detection policies for non-stationary regimes, such as the evening regime in section 4.

Acknowledgements The authors acknowledge the continuing support from the Defensive Information Warfare Branch at the Air Force Research Laboratory in Rome, NY. We are particularly grateful to W. John Maxey from AFRL, for his encouragement during the course of this effort. We are also indebted to BBN Technologies, especially to Steven Polin and Jinane Abounadi, for several discussions and support in the processing of the tcpdump data.

References

Figure 1. 1998 DARPA dataset - Wednesday of the 5th week: Number of connections under various time scales

(a) $T = 10\ s$

(b) $T = 100\ s$

(c) $T = 1000\ s$

(d) $T = 3600\ s$

Figure 2. ROCs induced by $X^T(k)$

(a) $X_r$

(b) $\rho$ population

Figure 3. ROCs induced by $X_r$ and $\rho$