Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.3, No.9, 2013

Probabilistic Models for Anomaly Detection Based on Usage of Network Traffic

Rohitha Goonatilake[1], Susantha Herath[2], and Ajantha Herath[3]
1. Department of Engineering, Mathematics, and Physics, Texas A&M International University, Laredo, TX 78041, USA
2. Department of Information Systems, St. Cloud State University, St. Cloud, MN 56301, USA
3. Administration and Technical Programs Division, University of Bahrain, Kingdom of Bahrain, Bahrain
* E-mail of the corresponding author: harag@tamiu.edu

Abstract
Recent advances in intrusions and attacks reflect vulnerabilities in computer networks. Innovative methods and tools can help attack defenses, prevent attack propagations, detect and respond to such attacks in a timely manner. Intrusion detection and prevention systems search for unauthorized use, recognize anomalous behavior, and prevent attempts to deny services. These systems gather and analyze information from the network, identify possible breaches of the security profile, as well as misuses. We have been experimenting with methods for introducing important concepts related to intrusion detection and improving undergraduate research experiences and education. To achieve this goal, probabilistic models are introduced to students in computer, information system and network security courses. This article presents a set of probabilistic methods and statistical models for network traffic anomaly detection. It also describes some prospects and how models have ripened from theories to big data analysis applications.
Keywords: Intrusion, conditional probability, network system, regression, data analysis

1. Introduction

Providing an extensive use of computers and network systems for legitimate activities, searching for unauthorized use, recognizing anomalous behavior, and preventing any attempts to deny users, machines, or portions of the network access to services are the tasks of intrusion detection systems (IDSs). The purpose of an Intrusion Detection System (IDS) is to gather and analyze information from various components within a computer or network, identifying possible breaches of the security profile, including unauthorized access as well as any other misuses. There are two types of IDSs: Network Intrusion Detection Systems (NIDS) and Host-Based Intrusion Detection Systems (HIDS). The monitoring approach is a primary factor in classifying the different types of IDSs. HIDS find suspicious activity or known attack patterns on the specific host where they occurred. NIDS gather data from the network traffic stream as it travels through the network components. Network traffic anomaly detection is to search and identify unauthorized activities of the network system. Potential users of these systems need to be aware including how well a given system is able to find intruders. This also includes determining the amount of work required to maintain a similar system in a fully-functioning network capacity allowing a significant volume of daily network traffic to occur. Researchers wanted to identify which prototypical attacks can be possible by the systems. Without accessing the nature of normal traffic generated by day-to-day work, they are unable to describe how well their systems could detect potential attacks while using background traffic information and avoiding false alarms. This information is critical as every intrusion requires time to review, regardless of whether it is a correct detection for which a real intrusion has occurred, or whether it is merely a false alarm as such the analyses of possible probabilistic and statistical models of variables employed.

This article begins with an example of network worm epidemics modeling. Section 3 describes the use of Bayes' theorem to outline scenarios under the systems accuracies that occur followed by a mean and standard deviation model. Point and interval estimations take the center stage describing possible techniques to estimate parameters of the models. The multivariate model will be broken down to include multiple regression, the regression model for scoring function, and data warehouse & model generator. The Markov process model and time series model will be given additional consideration thereafter. Lastly, conclusions and future work will provide a summary of the article together with possible undertakings as envisaged.

2. Network Modeling - Worm Epidemics

A simple example relates to a recent Federal criminal case in which documents provided by the prosecution counted the number of e-mail messages generated by a worm transmission. The payload of this worm was a 4kb packet. The network of the institution where the defendant worked could process packets at the rate of 1 Gb per second. To saturate the network, the worm would have to produce a minimum of 250,000 emails per second. According to the prosecution's data, only 261 e-mails were released into the network during the 3 hours the

worm was active. This number was not sufficient to degrade the performance of the network and the defendant was cleared. The spreading rate of this worm was hindered due to the strange file name given to the attachment, which was not attractive enough for many users to open. This knowledge can be expanded by examining epidemiological models.

Epidemics can be represented using a linear equation with a constant propagation rate, a non-linear equation with exponential growth rate, a differential equation or a difference equation. To obtain a useful prediction model, one should record the observations of all variables that may significantly affect the response to the epidemic. By virtue of its wide applicability, the linear model plays a prominent role in this process. Mathematical models express the laws that govern described actions and assumptions together with hypotheses relevant to the equations established within a study. For example, the uninfected computer components in a network under virus attack can be represented using (a linear equation or) a differential equation and the initial condition:

$$\begin{cases} \dfrac{dy(t)}{dt} = -ry(t),\ t > 0, \\ y(0) = y_0. \end{cases}$$

The general solution of this differential equation is $y(t) = y_0 e^{-rt}$, where the initial uninfected computer population is y_0 with a constant negative growth rate $-r$. The geometric representation of this general solution is an infinite family of integral curves, one similar to Figure 2-1.

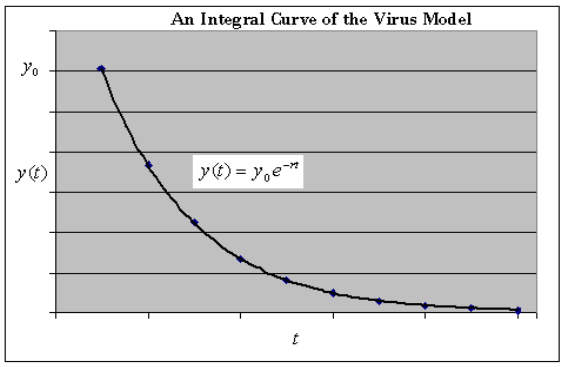

Figure 2-1. Decay of Un-Infected Computers Due to a Virus Epidemic

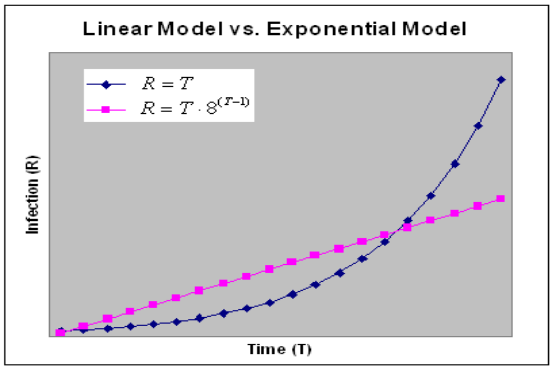

Figure 2-2. Growth of Infected Computers Due to Virus Propagations

A computer worm that randomly scans new hosts to attack and infect may follow the simple epidemiological model [1 & 2]. There are other applications that have been used in the past, stemming from pre-existing epidemiological, biological, and physical models. Figure 2-2 shows the number of infected computers over time due to two different types of attacks. At the initial detection stages, for smaller values of time T, the linear virus growth model, $R = T$, shows more infections than the exponential growth $R = T \cdot 8^{(T-1)}$. This false appearance leads us to make the wrong decision to allocate resources and assign a higher priority to suppressing the linear attack over the eventually more severe exponential attacks.

In many instances, there is a limit to the possible growth of epidemics. The logistic function models the restricted growth phenomenon as an exponential growth. As shown in Figure 2-3 below, the infection growth rate slows due to the limited capacity of the network, resulting in a logistic curve. The red curve to the left with the faster infection rate saturates the network sooner than the blue curve to the right. Figure 2-4 depicts the infection during an epidemic attack and the recovery due to the response. The growth of the number of infected computers begins in an exponential manner, levels out and eventually decays with the proper injection of response.

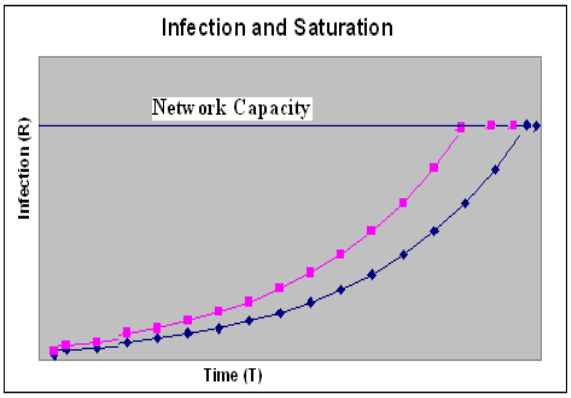

Figure 2-3. Infection and Saturation

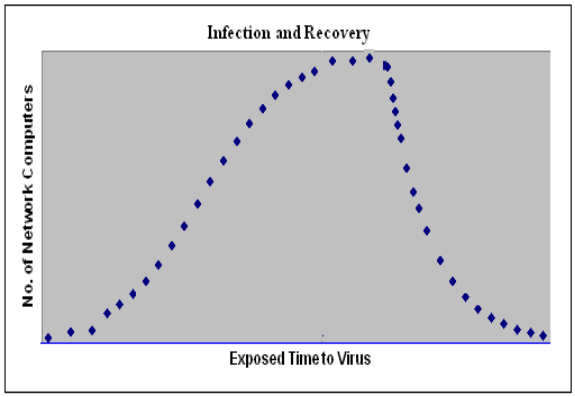

Figure 2-4. Infection and Recovery

Non-linear differential equations are used to model attack dynamics. The Epidemiological model derived from non-linear differential equations [1] can be represented by the equation:

$$n(t) = \frac{n_0(1-(d/\beta))}{n_0 + (1-(d/\beta) - n_0)e^{-(\beta-d)t}},$$

where $n(t)$ is the fraction of infected nodes, d is the infection or "death" rate, β is the recovery or "re-birth" rate, n_0 is the total number of vulnerable machines.

The Analytical Active Worm Propagation, AAWP, [6], gives a better estimate than Epidemiological modeling and can be mathematically represented as a difference equation:

$$n_{t+1} = sn_t + [N - n_t]\left[1 - (1 - 1/T)^{sn_t}\right],$$

where n_t is the fraction of infected nodes at time t, s is the scan rate, N is the total number of vulnerable machines and T is the address space the worm scans. The infection rate of the worm is proportional to the scanning rate. Random scanning has the searching ability to identify IP addresses to attack in a random order, localized scanning has the ability to attack targets that reside on the same subnet, hit list scanning has a list of targets to attack, permutation scanning uses a fixed pseudorandom permutation of IP addresses to attack and topological scanning uses the information stored in the victim's machine to attack the next target [1 & 2]. Worms that use localized scanning, like the Code Red II worm, spread at a slower rate than a worm that employs random scanning but has the capacity to penetrate firewalls.

3. Systems Accuracies

An IDS can generate four possible scenarios in an anomaly detection process. There are two types of incorrect detections an IDS can arrive at—falsely detecting actual intrusion and truly detecting false intrusion. Figure 3-1 below summarizes all possible scenarios when IDS is at work. A good IDS must reduce these false scenarios at any cost.

		Intrusion is:	
		Actual	False
Detection is:	Actual	IDS works	False alarm
	False	IDS did not work	IDS works

Figure 3-1. An Extent of IDS is at Work

From the theorem of total probability, and granted that $A_1, A_2, \ldots$ form a partition of the sample space, S, we have, for an event B, $P(B) = \sum_{i=1}^{\infty} P(B \mid A_i)P(A_i)$. It follows that

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{i=1}^{\infty} P(B \mid A_i)P(A_i)}$$ for any event A_i in the partition $A_1, A_2, \ldots$. This result is known as Bayes' theorem [3]. This presents the relationship between the conditional probabilities of various events involved in calculation of probabilities for common intrusion detection scenarios. Let us assume Table 3-1 of information for calculation of these probabilities.

Table 3-1. An IDS is at Work with Numbers

		Intrusion is:		
		Actual	False	Total
Detection is:	Actual	30	25	55
	False	23	32	55
	Total	53	57	110

Given that an IDS finds that there is an intrusion, what is the probability that it is in fact an intrusion? Note that there are only two events, A_1 and A_2 which correspond to the intrusion being an actual intrusion and a false alarm, respectively. The event B denotes that the intrusion is not being detected. Also, $A_2 = \overline{A}_1$. We want to evaluate $P(A_1 \mid B)$, the probability that an intrusion has taken place, given that the IDS finds it is a false alarm.

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} = \frac{0.482 \times 0.434}{0.482 \times 0.434 + 0.518 \times 0.561} = 0.419.$$

Many statistical models for intrusion detection stemmed from the operational, mean and standard deviation, multivariate (regression), Markov processes, and time series models [4]. Randomness is the basis for these models to exist; it perhaps represents one of the strangest challenges whose presence is successfully explored only by statistics tied to observable variables. The attempt of this section is to elaborate on these models more in-depth for an average reader. For a random variable X and n observations $X_1, X_2, \ldots, X_n$ the statistical model determines whether the subsequent observation X_{n+1} is a potential intrusion with respect to the previous n observations. The IDS provides tools to detect intrusion for the information assurance community based on the model under consideration. Knowledge-based intrusion systems collect knowledge about network attacks, examine network traffic and attempt to identify any patterns that indicate a suspicious activity has occurred. This merely applies against patterns of known attack and uses them to update the knowledge base frequently [5].

4. Mean and Standard Deviation Model

This operational model is primarily based on the assumption that intrusions can be decided by comparing a new observation of X against the previous n number of observations. This model can be applicable to metrics where experience shows that certain characteristics are frequently linked with attempted intrusions. An example is that an event counter is setup for the number of password failures during a brief period to suggest that a break-in has been successful. Another example is that personalities of the user such as the user first reads his e-mails, reply them before visit the file folders regardless of other network activities. Recovery of history sheds light on these types of anomaly detection for operational model. On the other hand, models based on statistical measures such as means and standard deviations provide a wide range of information for intrusion detection. It is often desirable to obtain an interval of values likely to contain the actual value of the unknown parameter. We then insist that the proposed interval contains the value with predetermined high probability called the level of confidence. This model is based on the assumption that all what we know about $X_1, X_2, \ldots, X_n$ are mean and standard deviation as given by $\overline{X} = \frac{\sum X_i}{n}$ and $s = \sqrt{\frac{\sum X_i^2 - \left(\sum X_i\right)^2 / n}{(n-1)}}$. A new observation X_{n+1} is

defined to be an intrusion if it falls outside a confidence interval $\overline{X} - t_{\alpha/2} \times \frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2} \times \frac{s}{\sqrt{n}}$ with a $(1-\alpha)\%$ degree of confidence. The symbol $t_{\alpha/2}$ denotes the $t-$value having an area of $\alpha/2$ to its right under the $t-$curve with degrees of freedom (df) equals $n-1$. When data are grouped in a frequency distribution, we use the computing formulas, $\overline{X} = \frac{\sum X_i \cdot f_i}{n}$ and $s = \sqrt{\frac{\sum X_i^{\,2} \cdot f_i - (\sum X_i \cdot f_i)^2 / n}{(n-1)}}$, where X_i denotes class midpoint, f_i denotes class frequency, and $n(=\sum f_i)$ denotes the sample size, to determine the mean and standard deviation for this confidence interval procedure. The basic assumptions for this procedure require that observations come from a normal population or a large sample and observations are gathered at random. Table 4-1 sums up the results of relevant parameter estimations.

Table 4-1. A Summary of Parameter Estimations

Assumption	Parameter to be Estimated	Confidence Interval at $(1001-\alpha)\%$ level
σ^2 known	μ	$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
σ^2 unknown	μ	$\overline{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$
μ unknown	σ^2	$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right)$

We now provide an example to show how these calculations are carried out using a relatively small data set. Other cases of parameter estimations cited follow similarly.

Suppose a signal is transmitted from a piece of network equipment to a receiver is normally distributed with mean μ and variance, $\sigma^2 = 4$. in order to avoid possible error, the same signal is sent 9 times repeatedly. These independent values are measured as 5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5. We now construct a 95% confidence interval for μ. Clearly, $\overline{X} = 81/9 = 9$. From Table 2, a 95% confidence interval for μ is $(7.69, 10.31)$. If the variance is unknown, then the sample variance, s^2, can be calculated from the sample data and estimations are calculated using the $t-$distribution. There can be instances in which the data points themselves contributed to outliers. They can be eliminated using the interval, $(Q_1 - 1.5(\mathrm{IQR}), Q_3 + 1.5(\mathrm{IQR}))$, where Q_1 and Q_3 be the quartiles and $\mathrm{IQR} = Q_3 - Q_1$. Other aspects of these calculations are found in any standard statistics text [6].

5. Point and Interval Estimations

Generally, estimations can be two-fold in this regard. In point estimation, we look for a single value for the unknown parameters. However, for the confidence interval estimations, a range of values are proposed giving lower and upper limits of the interval. The interval estimations can be based on the three-standard deviation rule, a given confidence level, or $\Gamma(.)$, the gamma function. The interval estimation based on $\Gamma(.)$, the gamma function is found in [7].

It is essential for forecasting future intrusions of the network, primarily from past data available from statistical forecasting techniques. The fact that the attacks occurred during a sequence of time periods are random variables. Let $X_1, X_2, \cdots$ be a sequence of random variables having the expected values, $E(X_1), E(X_2), \cdots$ which may or may not be independently and identically distributed, according to a given probability distribution. There are several potential forecasting procedures that can be used in the case of intrusions, which are listed below [8].

a) Last values: Let $\hat{E}(X_t)$ be the forecast for subsequent periods. For this estimation, the last value may be used as the possible evaluation in this method so that this estimator is $\hat{E}(X_t) = x_t$. Obviously, this estimating technique is disadvantageous because it yields a large variance as a result of selecting a sample of size 1.

b) Average: This average may be chosen to be $\hat{E}(X_t) = \sum_{i=1}^{t} \frac{x_i}{t}$ and can be a better estimator if the process is entirely stable. The use of a large amount of data can cause occasional shifts, when a reasonable estimation is desired.

c) Moving average: This estimation only uses the relevant recent data for the last periods, but is updated periodically placing more weight on x_{t-n+1} compared to x_t. Accordingly, the estimator is $\hat{E}(X_t) = \sum_{i=t-n+1}^{t} \frac{x_i}{n}$.

d) Exponential smoothing: This is, in fact, the weighted sum of the last observation and the one before using the smoothing constant, α, $0 < \alpha < 1$. As such, the recursive relationship $\hat{E}(X_t) = \alpha x_t + (1-\alpha)\hat{E}(X_{t-1})$ can be expressed alternatively as $\hat{E}(X_t) = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots$. Thus, the variance calculated to be, $\mathrm{Var}\left(\hat{E}(X_t)\right) = \frac{\sigma^2}{\frac{2-\alpha}{\alpha}}$. The choice of α determines the number of observations taken into consideration.

e) Exponential smoothing adjusted for trend: Let us assume that $E(X_t)$ follows a linear model with the slope S (a trend factor). The expected rate of findings can be either increasing or decreasing. The estimator is $\hat{E}(X_t) = \alpha x_t + (1-\alpha)\left[\hat{E}(X_{t-1}) + S\right]$. Since S is not generally known, it is estimated once again by using exponential smoothing to obtain $\hat{E}(S_t) = \beta\left[\hat{E}(X_t) - \hat{E}(X_{t-1})\right] + (1-\beta)\hat{E}(S_{t-1})$, where $0 < \beta < 1$ and $\beta \neq \alpha$ is becoming another smoothing constant. Accordingly, the new estimator of $E(X_t)$ is $\hat{E}(X_t) = \alpha x_t + (1-\alpha)\left[\hat{E}(X_{t-1}) + \hat{E}(S_{t-1})\right]$.

If X is a random variable with mean μ and variance σ^2, then Chebyshev's inequality states, for any value $k = d \times \sigma > 0$, $P(X :| X - \mu | \geq k)$, the probability of a value falling outside this interval is at most $1/d^2$. Note also that null occurrences should be included so as not to make the data biased. If X is the standard normal random variable and then, $E[X :| X - \mu |< t]$ is a strictly decreasing function for the nonnegative values of t [9]. This model is applicable to event counters, interval timers, and resource measures accumulated over a fixed time interval or between two related events. This model has several advantages over an operational model. The standard normal CDF table [10] facilitates computation of probability in order to decide if X_{n+1} has been an intrusion. Table 5-1 gives some crude estimates using three probabilistic principles in terms of the number of observations within a few standard deviations (s) measured from either side of the sample mean ($\bar{x}$).

Table 5-1. Estimates for the Number of Observations in a Data Set

Intervals Proposed	Three-Standard-Deviations Rule	The Empirical Rule	Chebychev's Rule $\left(1-\frac{1}{k^2}\right).100\%$
$(\bar{x}-s, \bar{x}+s)$	--†	68.26%	--
$(\bar{x}-2.s, \bar{x}+2.s)$	--	95.44%	75.00%
$(\bar{x}-3.s, \bar{x}+3.s)$	Almost all the observations	99.74%	88.89%

† no estimates are available

This model can be slightly modified placing greater weights on more recent values of the observations. For example, if $\omega_1, \omega_2, \ldots, \omega_n$ are positive weights such that $\omega_1 < \omega_2 < \ldots < \omega_n$ and $\sum_i \omega_i = 1$, then we consider the sequence, $\omega_1 X_1, \omega_2 X_2, \ldots, \omega_n X_n$ for the determination of mean and standard deviation of the

model.

6. Multivariate Model

Statistical modeling and data mining are two techniques that enable understanding and potentially identifying recurring events. This performance analysis examines trends and data considering the timeline of the event to determine statistically significant trends, both positive and negative, and the relationships among elements of an event using multivariate analysis. By establishing control or trend charts with confidence limits, these tools determine if an unusual behavior is present or has been attempted. This model is similar to the mean and standard deviation model except that it is based on correlations among two or more variable metrics. Experimental data shows that it is useful for better discriminating power obtained from combinations of related measures rather than individually and session elapsed time to look for any significant deviation of the trends [4].

6.1 Multiple Regression

Briefly, the purpose of multiple regression is to determine the relationship between the independent variable and the dependent variables. A more specific purpose of this technique is to investigate the importance of each independent variable to the dependent variable [11]. More specifically, this technique allows for the possibility to determine which predictor variable is most important, and which one is the least important. In this section, we extend the regression model to include more than one predictor variable. Thus, the general form of a linear equation with two independent variables is simply an extension of bivariate regression that can be written as $\hat{y} = a + b_1 x_1 + b_2 x_2$, where

$\hat{y}$ = the predicted value for the dependent variable y

a = intercept constant

b_1 and b_2 = regression coefficients (weights) for independent variables one and two

x_1 and x_2 = values for the two independent variables

In theory, the multiple regression equation could be extended to include an infinite number of predictors, but in practice (because of computational difficulties) it is perhaps prudent to limit one's analysis to a few variables. The multiple regression equation stipulates that the predicted value of y is a linear combination of an intercept constant, plus independent variables that are weighted by regression coefficients. Regression analysis yields the best-fitting values of coefficients using the least squares criterion that the sum of squared error terms (the differences between y and predicted value of $\hat{y}$) are kept at minimum. This technique is often referred to as the least squares-method subjected for much of the analysis.

6.2 Regression Model for Scoring Function

NETADs (Network Traffic Anomaly Detectors), such as PHADs (Packet Header Anomaly Detectors), detect anomalies in network packets [12]. A possible statistical model for quantities involved in the evaluation of NETAD anomaly scores for a packet; this can be expressed by $\text{NETAD} = \sum tn_\alpha (1 - r/256)/r + t_i/(f_i + r/256)$, where r is the size of values (up to 256 for NETAD), t is the time since the attribute was last anomalous (during either training or testing), n_α is the number of training packets from the last anomaly to the end of training period, and f_i is the frequency in training. The summation is taken over all possible subset/attribute combinations. Table 6-1 shows the number of attacks detected from 20 to 5000 false alarms using various anomaly scoring functions as appeared in [13].

Table 6-1. Factors Related to Scoring Function [10]

Scoring Function	tn/r	tn_α/r	$tn_\alpha(1-r/256)/r$	$t_i/(f_i+1)$	$t_i/(f_i+r/256)$	**NETAD**
20	56	56	60	33	78	66
50	78	89	92	52	115	97
100	104	118	120	81	127	132
500	141	148	149	130	142	148
5000	157	152	152	158	156	152

The relationship between the scoring functions and tn/r is derived for values in Table 6-1.

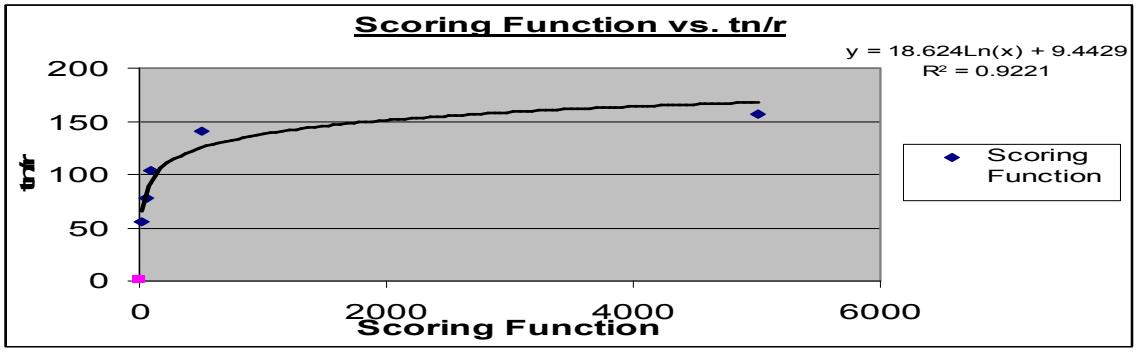

Figure 6-1. Graph of Scoring Function vs. tn/r

This shows that the scoring function satisfies the relation $tn/r = 18.624 \times \ln(\text{Scoring Function}) + 9.4429$ with a strong correlation coefficient. It is evident that the increase of scoring function beyond a certain value will not increase the tn/α so rapidly. That is, tn/α slows down as it reaches higher values of the scoring function as in Figure 6-1. In many situations, the response variable is not a linear function of the input function. In such cases, the determination of a relationship is possible by a change of variables to transform it into a linear form.

6.3 Data Warehouse and Model Generator

For misuse of detection problems, the Support Vector Machine (SVM) model classifies the network activity as normal, or as belonging to one of the four categories of attack, namely, probe, dos, u2r, and r2l, where the letter "a" in parentheses is used to indicate if they are actual, and "p" is used to indicate the predicted values. The misuse classification results are summarized in Table 6-2 for detection dataset obtained [14].

Table 6-2. Confusion Matrix on DARPA Intrusion

Actual (a)/ Predicted (p)	normal (p)	probe (p)	dos (p)	u2r (p)	r2l (p)
normal (a)	59332	1048	45	57	111
probe (a)	602	3251	212	62	39
dos (a)	7393	88	222288	75	9
u2r (a)	178	1	8	33	8
r2l (a)	14683	41	7	31	1427

Table 6-3. Correlation Matrix (Actual vs. Predicted)

Category	normal (p)	probe (p)	dos (p)	u2r (p)	r2l (p)
normal (a)	1				
probe (a)	-0.05389	1			
dos (a)	-0.20494	-0.31953	1		
u2r (a)	0.102886	0.382839	0.686313	1	
r2l (a)	0.023286	-0.32087	-0.27926	-0.59802	1

Table 6-3 is the correlation matrix for actual and predicted misuses of detection for the data summary of independent and dependent variables. The most important aspect to notice from this table is that no bivariate correlation is greater than 0.686. We are safe to conclude that the analysis is free of any possible multicollinearity effects. Those interested to find out more about SPSS features to test for multicollinearity can refer to [9]. A negatively correlated pair of items in a detection dataset indicates that if actual values increase, then predicted values decrease and vice versa [15].

The detection rate and false positive rate have been obtained for some varying thresholds for the determination of the system's accuracy [16]. The detection rate is the percentage of attack records correctly identified. The false positive rate is the percentage of normal records misdiagnosed as anomalous. The threshold is the value that determines if a record is normal or if it is a possible attack. Table 6-4 below includes a sample of the varying threshold and corresponding false positive rate and the detection rate. We establish confidence intervals for threshold, false positive rate, or detection rates and other descriptive statistics with a 95% degree of confidence.

Table 6-4. A Sample of Varying Threshold vs. False Positive and Detection Rates [16]

Threshold	False Positive Rate (%)	Detection Rate (%)
-1.08307	0.790142	0.373533
-1.08233	0.828005	0.480256
-1.07139	1.54441	0.533618
-0.968913	1.65734	1.17396
-0.798767	3.58736	3.89541
-0.79858	3.63784	5.60299
-0.798347	3.68999	6.77695
-0.767411	3.72054	6.83031
-0.746663	4.35691	7.47065
-0.746616	4.63025	8.00427
-0.71255	8.34283	20.9712
-0.712503	8.75201	22.0918

Regression models have been obtained for these variables [17]. We draw attention to the descriptive statistics that are naturally a focus in every preliminary study for these three variables in Table 6-5. A correlation matrix provides significant evidence to conclude that there are strong correlations among the variables in the study as summarized in Table 6-6.

Table 6-5. Descriptive Statistics for Threshold, False Positive, and Detection Rate

	Threshold	False Positive Rate (%)	Detection Rate (%)
Mean	-0.85726	3.79480	7.01708
Standard Error	0.04306	0.74782	2.12850
Median	-0.79846	3.66392	6.18997
Standard Deviation	0.14916	2.59053	7.37333
Sample Variance	0.02225	6.71086	54.36600
Kurtosis	-1.26349	0.29767	1.10407
Skewness	-0.78960	0.89255	1.37868
Range	0.37057	7.96187	21.71827
Minimum	-1.08307	0.79014	0.37353
Maximum	-0.71250	8.75201	22.09180
Sum	-10.28714	45.53763	84.20495
Count	12	12	12
Largest(1)	-0.71250	8.75201	22.09180
Smallest(1)	-1.08307	0.79014	0.37353
Confidence Level (95.0%)	0.09477	1.64595	4.68479

Table 6-6. Correlation Matrix for Threshold, False Positive, and Detection Rate

	Threshold	False Positive Rate (%)	Detection Rate (%)
Threshold	1		
False Positive Rate (%)	0.838230425	1	
Detection Rate (%)	0.750245026	0.983370245	1

False alarm and detection rates over 50 users and the tradeoff between false alarms and detection ability as functions of the offset have also been considered [18]. Control charts can be generated under simulated scenarios to provide explicit details of a charting procedure for future work and draw any long-term conclusions that determine the relationship between the variables.

7. Markov Process Model

A probabilistic prediction model identifies the last system identification given all the previous system identifications in the detection of the traffic anomaly are known. From this, the prediction model is same as finding probability, $P[X_{n+1} \mid X_1, X_2, \ldots, X_n]$ [19]. Let $\{X_i : 1 \le i \le N\}$ denote the traffic measurement in time series and let W denote the number of samples for the averaging process. The given trace $\{X_i\}$ is now

extended to a series of pairs $\{(X_i, \overline{X_i})\}$ with $\overline{X_i} = \frac{1}{W}\sum_{k=i-W}^{i-1} X_k$ for $i = W+1,\ldots,N$. For a given time series with fast decaying autocorrelation function, the average values $\overline{X_i}$ will be rather good estimators of the sample mean and will be almost the same for all values of i [20]. Hence, this additional category contains no information and is of very limited value with respect to a better characterization of the sample correlations. However, in the presence of a slowly decaying autocorrelation function, the values $\overline{X_i}$ are significant and represent the cumulated averaged history of $\overline{X_i}$. The series of pairs consists of W samples less than the original time series [18]. This model, which applies only to event counters, regards each distinct type of event (audit record) as a state variable, and uses a state transition matrix to characterize the transition frequencies between states. A new observation is defined to be an intrusion if its probability as determined by the previous state and the transition matrix is too low. This model might be useful for looking at transitions between certain commands where command sequences were important [21]. Whenever the intrusion attempt is to occur in state i, there is a fixed probability p_{ij} that will be in state j. Let E be a finite or countable set. Let $P = (p_{ij})$ be a stochastic $E \times E$ matrix, so that, for $i, j \in E$, we have $p_{ij} \geq 0, \sum_{k \in E} p_{ik} = 1$. Let μ be a probability measure on E [22]. We know that there exists a triple (Ω, F, P^{μ}) carrying a Markov chain $Z = (Z_n : n \in Z^{+})$ such that $P(Z_0 = i_0; Z_1 = i_1; Z_2 = i_2; \ldots, Z_n = i_n) = \mu_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}$ holds. We write 'a.s., P^{μ}' to signify 'almost surely relative to the $P^{\mu}-$ measure'. Let $F_n := \sigma(Z_0, Z_1, \ldots, Z_n)$, then (a.s., P^{μ}), $P^{\mu}(Z_{n+1} = j \mid F_n) = p_{Z_n j}$, where F_n is the natural filtration [23].

8. Time Series Model

The time series model assumes no conditions in the analysis. It compares the parameters collected over an extended period of time with the current parameters to monitor unusual features, namely, trends, seasonal effects, cycles, and irregularities of the prevailing data.

This model uses an interval timer together with an event counter or resource measure, that takes into account the order and inter-arrival times of the observations $X_1, X_2, \ldots, X_n$, as well as their values. A new observation considers an attempted intrusion if its probability of occurring at that time is considerably small. A time series model has the advantage of measuring trends of behavior over time and detecting gradual but significant shifts in behavior, but the disadvantage is that it is more costly compared to other existing models. Furthermore, when the parameters describing the series do not change over time, the time series sometimes can be modeled adequately by using what is called the multiplicative decomposition model that exhibits trends, seasonal effects, and other features of the series [24].

The multiplicative decomposition model is $y_t = TR_t \times SN_t \times CL_t \times IR_t$, where y_t is the observed value of the time series in time period t and other components (factors) represent trends, seasonal, cyclic, and irregular components in time period t, respectively. First, the time series is plotted using any available data for a given network. A cyclical component refers to repeating up and down movements around trend levels (for example that revolve around the business cycle.)

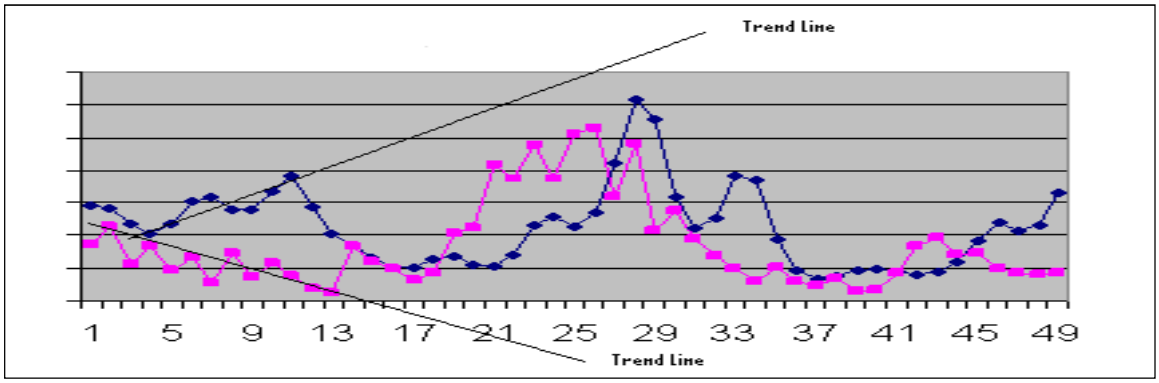

Figure 8-1. Trends, Seasonalties, Cyclicties and Irregularities

These fluctuations can last anywhere from smaller to longer periods as measured from peak to peak or trough to trough. The two trend lines drawn in Figure 8-1 differ substantially to indicate that there has been heavy traffic volume during that time period. Seasonality has not been a question. Cyclic features remain the same, whereas some irregularities are present at the beginning and end of the time period monitored.

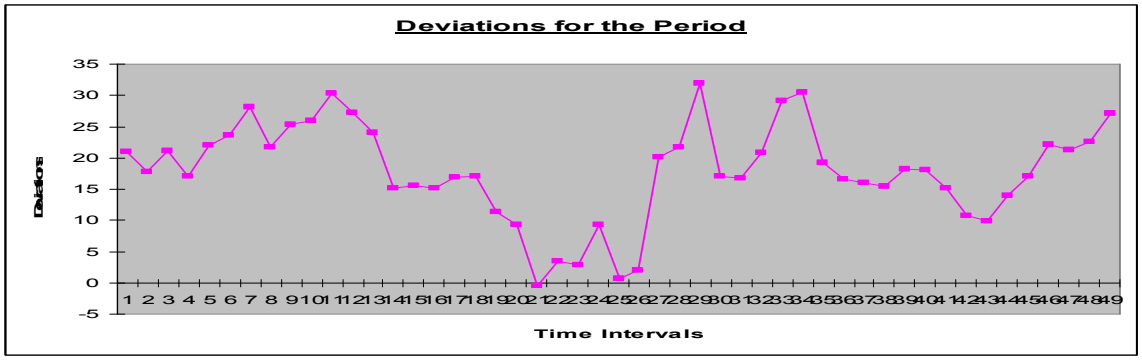

Figure 8-2. Deviation Plot from Comparison

The deviations have been computed for all values of time variable. This has resulted in the deviation plot appears in Figure 8-2. The extremes reflect that potential anomalies may have occurred.

Another tool available in the arsenal is the chi-square goodness-of-fit test. It is a hypothesis test which can be used to make inferences about the distribution of a variable from the information collected from a sample data to decide whether the distribution in question differs from that of a known population. A sample of user activities on a suspected computer network is analyzed to determine whether activity pattern differs from the activities previously recorded in the network. The network activity patterns in a computer system are recoded through a stream of audit events to make this determination. One advantage of this procedure is that it is a robust test [21].

9. Conclusions and Future Work

Citing brief notes from several articles and presenting a set of statistical models followed by some analytical discussions, we were able to present a set of probabilistic methods and statistical models for network traffic anomaly detection. A more comprehensive procedure and formalism for the determination of the normal profile is always necessary. It is important to investigate whether or not other scenarios may have occurred. A decent survey in which the scope of current research extensively documented is provided in [25]. Discretization methods convert the numeric time series to symbolic time series for the method of additional knowledge discovery [26 & 27]. However, at present, analyses on multivariate data rather than univariate data are performed to detect anomalies as they do not have a strong signature in any of the time series of individual features [28]. Other statistical models of this type are there to address the anomaly detection, for example, models that use more than the first two moments (e.g., $E(X), E(X^2)$) but less than the full set of values. If the determination is done on the basis of function of random variables, the method of statistical differentials can be used to evaluate the first two moments as discussed [29]. Parameter estimation is the key in some of the models. Much focus has been given to estimation based on confidence interval. The maximum likelihood methods and the

Bayesian approach are two other well-known parameter estimation methods that can be used as well. Simulation has been used in the literature for determination of parameters for these models. However, one should also note that the limitations and difficulties of first generation of attacks could prompt the attacker to search for new techniques to devise and execute more harmful attacks to computer network [30].

Acknowledgements
The authors appreciate the work of Gladys Gonzalez, Sofia C. Maldonado, and Navil Lozano for reading the many versions of this article for improvement.

References

1. Chen, Zesheng, Gao, Lixin, Kwiat, Kevin, *Modeling the Spread of Active Worms*, IEEE Infocom, http://www.ieee-infocom.org/2003/papers/46_03.PDF, 2003
2. Ellis, Daniel R., Aiken, John G., Attwood, Kira S., Tenaglia, Scott. D., *A Behavioral Approach to Worm Detection*, Proceedings of ACM Workshop on Rapid Malcode, 2004
3. Harold J. Larson, *Introduction to Probability*, Addison-Wesley Advanced Series in Statistics, Addison-Wesley Publishing Company, Reading, MA, 1995
4. Dorothy E. Denning, *An Intrusion-Detection Model*, IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, February 1987, pp. 222—232
5. H.-Y. Chang, S.F. Wu, and Y.F. Jou, *Real-Time Protocol Analysis for Detecting Link-State Routing Protocol Attacks,* ACM Trans. Inf. Sys. Sec., Vol. 1, 2001, pp. 1-36
6. Sheldon M. Ross, *Introduction to Probability and Statistics for Engineers and Scientist,* third edition, Elsevier Academic Press, San Diego, CA, 2004
7. Rohitha Goonatilake, Rafic Bachnak, and Susantha Herath, *Statistical Quality Control Approaches to Network Intrusion Detection,* International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011, pp. 115-124
8. Frederick S. Hillier and Gerald J. Lieberman, *Operations Research*, Second Edition, Holden-Day, Inc., San Francisco, CA, 1967
9. Stephen J. Herschkorn et al., *Decreasing Expectations,* Mathematics Magazine, Vol. 79, No. 2, pp. 155
10. *Cumulative Normal Probability Tables (Z-Values),* http://www.osat.umich.edu/sixsigma/Reference/norm-tables.PDF
11. B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, Fourth Edition, Allyn & Bacon, Needham, Heights, MA, 2001
12. Matthew V. Mahoney, *Network Traffic Anomaly Detection Based on Packet Bytes*, The Eighteenth Annual ACM Symposium on Applied Computing (SAC 2003), March 9 - 12, 2003, Florida Institute of Technology, Melbourne, FL, 2003 http://www.cs.fit.edu/~mmahoney/paper6.pdf
13. A. Laxmi Kanth, Suresh Yadav, and M.Sridhar, *Hybrid Modular Approach for Anomaly Detection*, International Journal of Computer Applications (0975 – 8887), Vol. 14, No.8, February 2011
14. Marcos M. Campos and Boriana L. Milenova, *Creation and Deployment of Data Mining-Based Intrusion Detection Systems in Oracle Database 10g*, Oracle Data Mining Technologies, 2005, http://www.oracle.com/technology/products/ bi/odm/pdf/odm_based_intrusion_detection_paper_1205.pdf
15. *Results of the KDD'99 Classifier Learning,* http://www.cse.ucsd.edu/users/elkan/clresults.html, 1999
16. Katherine A. Heller, Krysta M. Svore, Angelos D. Keromytis, and Salvatore J. Stolfo *One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses,* Columbia University, http://www1.cs.columbia.edu/~kmsvore/ocsvm.pdf
17. Ajantha Herath et al., *Mathematical Modeling of Cyber Attacks: A Learning Module to Enhance Undergraduate Security Curricula*, The Journal of Computing Sciences in Colleges (JCSC), Vol. 22, Is. 4, 2007, pp. 152-161
18. William DuMouchel, *Computer Intrusion Detection Based on Bayes Factors*
For Comparing Command Transition Probabilities, National Institute of Statistical Sciences (NISS), Technical Report, No. 91, February, 1999
19. Eleazar Eskin, Wenke Lee, and Salvatore J. Stolfo, *Modeling System Calls for Intrusion Detection with Dynamic Window Sizes*, Computer Science Department,
Columbia University and Computer Science Department, North Carolina State University, http://www1.cs.columbia.edu/ids/ publications/smt-syscall-discex01.pdf
20. O. Rose, *A Memory Markov Chain Model for VBR Traffic with Strong Positive Correlations*, Institute of Computer Science, University of Würzburg, http://www.iai.inf.tu-dresden.de/ms/rose/papers/1999itc.pdf
21. Rohitha Goonatilake et al., *Intrusion Detection Using the Chi-Square Goodness-of-Fit Test for Information Assurance*, Network, Forensics and Software Security, The Journal of Computing Sciences in Colleges

(JCSC), Vol. 23, No. 1, 2007, pp. 255-263
22. Nong Ye, Xiangyang Li, Qiang Chen, Syed Masum Emran, and Mingming Xu, *Probabilistic Techniques for Intrusion Detection Based on Computer Audit Data,* IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 31, No. 4, July 2001
23. David Williams, *Probability with Martingales*, Cambridge University Press, 1991
24. Bruce L. Bowerman and Richard T. O'Connell, *Forecasting and Time Series: An Applied Approach,* Third Edition, Duxbury Thomson Learning, 1993
25. Varun Chandola, Arindam Banerjee, and Vipin Kumar*, Anomaly Detection: A Survey*, ACM Computing Surveys, Vol. 41, No. 3, Article 15, July 2009, pp. 15-58
26. Fabian Mörchen and Alfred Ultsch, *Optimizing Time Series Discretization for Knowledge Discovery,* KDD'05, August 21-24, 2005, Chicago, IL.
27. Jerome L. Myers and Arnold D. Well, *Research Design and Statistical Analysis,* Second Edition, Lawrence Erlbaum Associates, Inc. Publishers, 2003
28. Jeff Terrell, Kevin Jeffay, F. Donelson Smith, et al., *Multivariate SVD Analyses For Network Anomaly Detection,* University of North Carolina at Chapel Hill,2005, http://www.cs.unc.edu/~jeffay/papers/SIGCOMM-05.pdf
29. Rohitha Goonatilake, *On Method of Statistical Differentials,* African Diaspora Journal of Mathematics (ADJM), Vol. 3, No. 2, 2005, pp. 25—52
30. V. Anil Kumar, *Sophisticated in Distributed Denial-of-Service Attacks on the Internet*, Current Science, Vol. 87, No. 7, 2004, pp. 885-888