2010 Second International Conference on Communication Software and Networks

ARIMA Based Network Anomaly Detection

Asrul H. Yaacob, Ian K. T. Tan
FIST, FIT
Multimedia University
Cyberjaya 63100 Malaysia
{asrulhadi.yaacob,ian}@mmu.edu.my

Su Fong Chien
Dept. of Physics, Faculty of Science
University Malaya
Petaling Jaya 50603 Malaysia
sfchien@um.edu.my

Hon Khi Tan
Global Service Delivery
Free Net Business Solutions Sdn Bhd
Cyberjaya 63000 Malaysia
sharyn@fnbs.net

Abstract— **An early warning system on potential attacks from networks will enable network administrators or even automated network management software to take preventive measures. This is needed as we move towards maximizing the utilization of the network with new paradigms such as Web Services and Software As A Service. This paper introduces a novel approach through using Auto-Regressive Integrated Moving Average (ARIMA) technique to detect potential attacks that may occur in the network. The solution is able to provide feedback through its predictive capabilities and hence provide an early warning system. With the affirmative results, this technique can serve beyond the detection of Denial of Service (DoS) and with sufficient development; an automated defensive solution can be achieved.**

Keywords-Intrusion Detection, Denial of Service, ARIMA, Forecasting, Network Security

I. INTRODUCTION

With the Internet becoming ubiquitous, the importance of preventing malicious activities from happening on the network is unprecedented. Network administrators would rather take preventive measures than fix problems that may involve days of system downtime that organizations may not be able to afford. The network administrator is to ensure that systems and end-users are protected from malicious individuals who continuously launch attacks in a variety of forms. Such violations of network security are deemed as network intrusion. Generally, they are two proposals to handle or avoid the violations mentioned above; (1) intrusion prevention and (2) intrusion detection. The former uses identity authentication that ranges from the validation of usernames and passwords to the use of biometrics. The latter, which is the focus of this paper, utilizes complex and intelligent computational model in detecting an attack. With an early warning system that is able to alert the network administrator just a few seconds before a malicious attack on the network, the network can be shutdown and traffic can be re-routed to a secondary or backup network connection.

One of the key issues in network traffic is to able to forecast the future traffic rate based on measured traffic history. Network traffic predictability denotes the possibility for prediction to satisfy some precision requirement over desired prediction/control time interval. In respect to that, one of the most important and widely used time series models is the Autoregressive Integrated Moving Average (ARIMA) models. The popularity of the ARIMA model is due to the statistical properties as well as the well-known Box-Jenkins methodology in the model building process.

Intrusion detection techniques can be categorized into signature detection and anomaly detection [1][2][3]. Signature detection systems identify intrusions based on the experiences from previously known patterns of attacks or weak areas in the system. An alarm signal is generated when the captured traffic matches with the attack pattern. However, the disadvantage of this technique is that it has no ability to recognize new attack formats. Anomaly detection systems will signal anomalies occurrence when the observed activities consists of large deviation from the established normal usage. This system will generally employ a probability model to determine whether an activity can be deemed as abnormal. The merit of this technique is that it does not need to gain prior experience to detect attacks and hence it can provide an alert of potential attacks. However, the deficiency of this technique is that it may not be able to provide detailed attack information and even may produce high false positive rate.

This paper develops a novel idea by incorporating ARIMA for network-based intrusion detection system that is based on anomaly network traffic detection technique. By using ARIMA, we can predict the future trend of a normal traffic based on the previous data. This will give us a progressive and real-time estimation of the normality for a system. Knowing the expected network traffic pattern, we can match real-time network traffic with the forecasted network traffic and upon detecting a specific difference over a preset threshold, an alert can be raised.

II. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

In the IMSL C Numerical Library time series modules, there are three methods which have been developed for the ARIMA models. The theory of the ARIMA can be written as below [4][5][6]:

A small, yet comprehensive, class of stationary time-series model consists of the non-seasonal ARMA processes is defined by

$$\phi(B)(\ y_t\text{-}\mu)=\theta(B)\ A_t,\ \ t\in Z \qquad (1)$$

978-0-7695-3961-4/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCSN.2010.55

where Z denotes the set of integers, B is the backward shift operator defined by $B^k y_t = y_{t-k}$, μ is the mean of Y_t, and from equation (2), we may obtain

$$\phi(B) = 1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_P B^P,\ p \geq 0 \quad (2)$$

$$\theta(B) = 1-\theta_1 B - \theta_2 B^2 - \ldots - \phi_{2q} B^q,\ q \geq 0 \quad (3)$$

where p is the number of autoregressive parameters and q is the number of Moving average parameters. Equations (1)-(3) form a model so-called ARMA(p,q) of order (p,q) [7].

If the "raw" data, { y_t }, are homogenous and non-stationary then they can be differentiated to produce a new set of stationary data, and this model is referred to as ARIMA. Parameter estimation is performed on the stationary time series $y_t = \nabla^d y_t$, where $\nabla^d = (1\text{-}B)^d$ is the backward difference operator with period 1 and order d, d > 0. If the data consists of seasonal trend then the advanced and sophisticated model ARIMA (p, 0, q) × (0, d, 0)$_s$ is applied, and this model can be written as

$$\phi(B) = \nabla^d_s(y_t\text{-}\mu)= \theta(B)a_t,\ t=1,2,\ldots,n. \quad (4)$$

where s is the seasonal fit parameter, $\nabla^d_s=(1\text{-} B^s)^d$, Z_t is unobserved outliers-free data with mean μ, and a_t is the associated white noise . It is assumed that all roots of ϕ(B) and θ(B) lie outside the unit circle and when s = 1 equation (4) reduces to the conventional ARIMA(p,d,q).

It is noted that IMSL ARIMA model does not treat the data as observable but rather than the values that may contain one or more outliers. Since p, q, s, and d are optional parameters, IMSL has developed three methods as follows:

A. *Method 1 (M1): Automatic ARIMA(p,0,0) x (0,d,0)$_s$ Selection*

This method initially searches for the AR(p) representation with minimum Akaike's An information criterion (**AIC**) for the noisy data, where p = 0,..., maximum number of AR parameters requested. This method ensures every possible combination of values for p, s, and d is examined. If s = 1 and d = 0, this leads pure autoregressive prediction.

B. *Method 2 (M2): Grid Search*

The second automatic method conducts a grid search for a set value of p and a set value for q. Grid search can be extended to include the candidate values for s and d. This method does a thorough search for all possible combinations in order to obtain minimum AIC. However, this consumes lots of time in predicting future data.

C. *Method 3 (M3): Specified ARIMA (p, 0, q) ×(0, d, 0)$_s$ Model*

In the third method, specific values for p, q, s and d are given. If the set values of s and d are defined, then a grid search for the optimum values of s and d is conducted.

With the possibility to predict the future pattern, ARIMA could be used to create a progressive normal model of a system. Projection based on a "normal" data will result in a "normal" data. This prediction can be used as the model of normality in anomaly-based IDS. The model created follows the latest pattern of the system, thus making it very dynamic.

In defining the model for the system, different parameters were used. Several models were defined in order to detect different types of anomaly. This is largely due to different categories of attacks that exist today. As for example, UDP flooding attack is largely due to lack of flow control in UDP, whereas SYN flooding target the weakness in TCP handshaking process. A study indicated that more than 90% of DoS attacks use TCP as their protocol of choice [3]. We will focus only in these two types of attack, which can be identified as DoS.

In the first type of attack, UDP flooding, the attack will send as many as possible UDP datagrams to the target. Due to the absence of flow control in UDP, the attacker will certainly consume not only the bandwidth of the target but also the CPU and memory utilization. Even though the target will simply discard the datagram, this will cause a denial of service for the target. In order to detect this kind of attack, ARIMA could be used to predict the "normal" traffic pattern. By measuring the volume of traffic, incoming and outgoing, we could define the model of normality.

In the second type of attack, TCP SYN flooding, we will see an increased number of incomplete TCP handshaking process. Even though the IP in the first SYN packet is spoofed, the target will be a real machine. The target will simply wait for the ACK, which will never come, to complete the handshaking process. Thus, by measuring the number of SYN packets as well as the number of completed TCP handshaking process, it will provide an indication of a possible attack. In normal conditions, the number of SYN packets will certainly match the number of completed handshaking process. We could use ARIMA to predict a correct ratio of SYN packets to the number of completed TCP handshaking.

The prediction made using ARIMA is considered as the expected forecasted normal data. Several studies have demonstrated ARIMA could predict very well, and base on the argument of any deviation over a provided threshold from predicted value will be considered as an anomaly. The network traffic is monitored to get statistics on the usage. Based on this, a predicted traffic volume is computed using ARIMA. A threshold is set to determine whether the real traffic volume follows the forecasted normal traffic volume.

III. Implementation

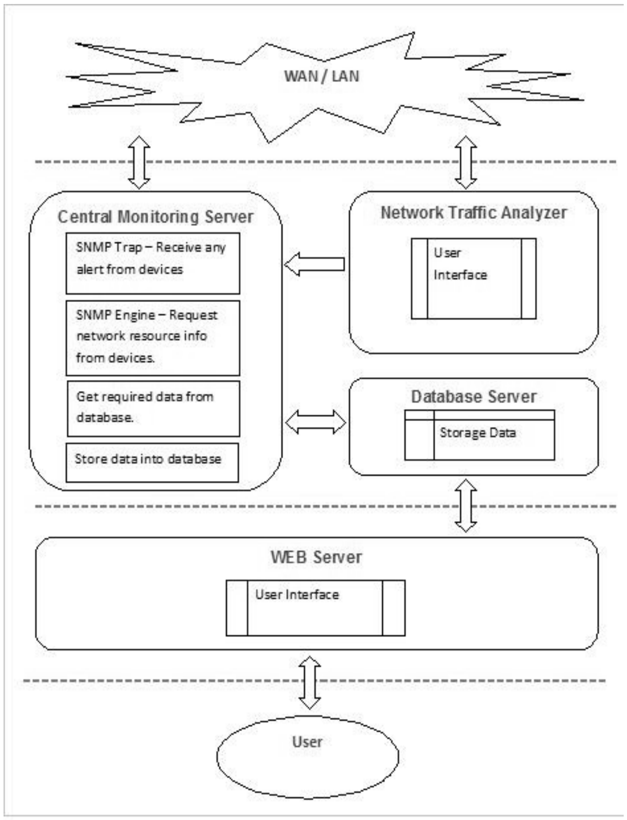

WAN / LAN
Central Monitoring Server
SNMP Trap – Receive any alert from devices
SNMP Engine – Request network resource info from devices.
Get required data from database.
Store data into database
Network Traffic Analyzer
User Interface
Database Server
Storage Data
WEB Server
User Interface
User

Figure 1. Overall System Architecture.

Figure 1 depicts the overall system implemented. It comprises of three main components; the Central Monitoring Server (CMS), the Database Server (DBS) and the Network Traffic Analyzer (NTA). The Central Monitoring Server traps Simple Network Management Protocol (SNMP), logs it into the Database server and through a web interface displays the messages. The CMS is written in C# and runs on Microsoft Windows. The database used is MySQL 3.23. The NTA, which is the main component of this paper is coded in C on the GNU Linux environment.

The IMSL Numerical Library ARIMA methods were used to provide the prediction. This library consists of a comprehensive set of mathematical and statistical functions. The functionalities of this system are:

i. Capturing the traffic in real time
ii. Creating statistics based on the captured traffic
iii. Predicting the normal pattern for the current traffic
iv. Measure the difference between the predicted traffic and real traffic
v. Signal to the CMS any abnormality detected

Figure 2 shows the flow of the NTA program. There will be a main process that controls the real time capture of the network packets. Other processes include processing of the received packets, generating of the statistics and computing the forecasted traffic. It is noted that the processing and statistics modules are actually executed independently from the main program.

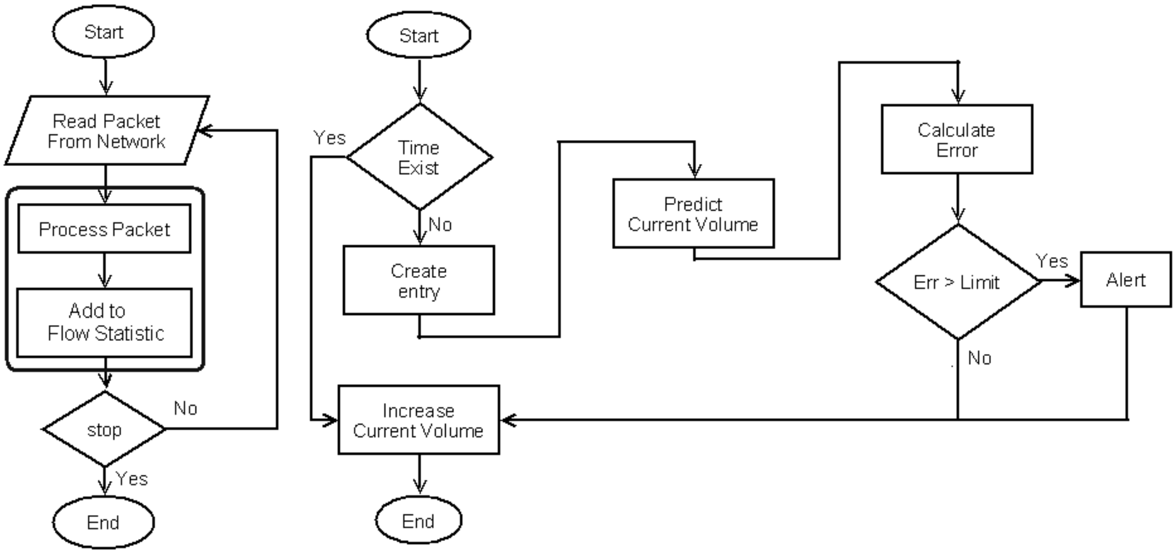
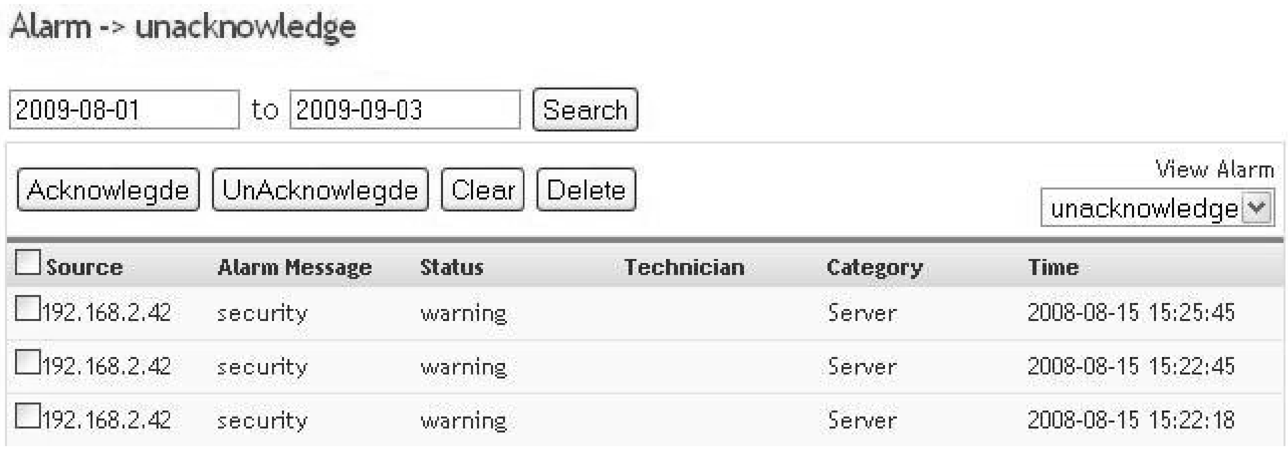


Figure 2. NTA Program Flow.



2009-08-01 to 2009-09-03

□ Source	Alarm Message	Status	Technician	Category	Time
□ 192.168.2.42	security	warning		Server	2008-08-15 15:25:45
□ 192.168.2.42	security	warning		Server	2008-08-15 15:22:45
□ 192.168.2.42	security	warning		Server	2008-08-15 15:22:18

Figure 3. Capture SNMP on CMS from NTA.

In the statistics module, the volume of traffic is collected and is used for prediction. The data for the statistics module is generated every second. The traffic is then discarded once calculation is done. The collected traffic will be used to predict the expected normal pattern only. In this implementation, the traffic pattern prediction is based on the previous 20 seconds of data by employing the ARIMA algorithm. Since the function of ARIMA is computationally intensive, one module is created to handle the prediction. The predicted results will then return to the statistics module after each computational cycle. The statistics module will then proceed to compare the real traffic pattern with the predicted results. The difference between them will be used to decide the abnormality of the traffic. An alert via SNMP will be sent to the CMS (Figure 3) if the difference exceeds a threshold value.

IV. RESULTS AND DISCUSSION

The program runs on a standalone workstation, it analyses real time packets that were captured in a live data center. The captured real traffic volume, predicted traffic volume and the computed errors for the prediction are depicted in Figure 4. As indicated previously, ARIMA methods are used to predict the expected normal traffic pattern. From the computational experiment, we found that Method 3 has the best fit for our empirical studies since it takes the least time to complete one cycle of prediction and its accuracy is within tolerance. Hence, the results presented here are based on method 3.

If the pattern is stable enough, the predicted traffic should match the normal traffic well; in our case we set the prediction error to be less than 15%. We could also see the false positive alarm generated at time 53s mainly due to low real traffic at that time. Error rates are computed by using the absolute value derived from the division of the variation between the forecasted traffic volume and the actual traffic volume against the actual traffic volume.

As such during low actual traffic volume conditions the error rates will usually be higher as compared to high traffic volume conditions. Suitable error rates threshold will need to be determined depending on the network environment. This would imply that it is not a suitable technique for small network environment where traffic volume is generally low with many burst of network activities.

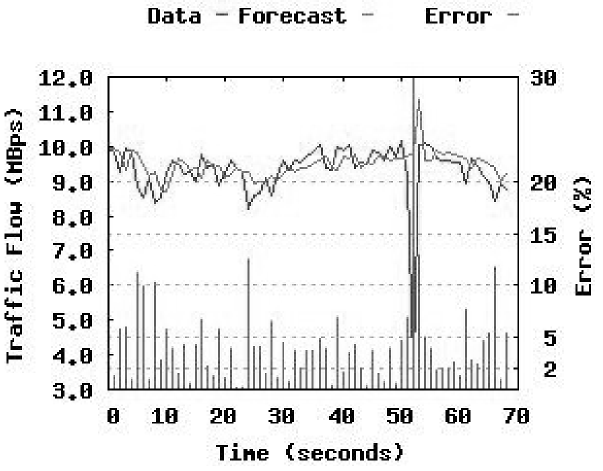


Figure 4. Output during normal traffic condition with one false positive.

The program will alert when the error in prediction exceeds the threshold. From our empirical studies of our data center environment, we found that a threshold of 15% is suitable. If it exceeds this threshold, it indicates a potential attack attempt such as UDP flood. This is evident in Figure 5. As can be seen from the figure, there is a sudden increase in volume of traffic which might suggest a DoS attack. Based on this observation, an anomaly on the traffic can be detected by comparing a real traffic volume to the predicted volume that was generated by ARIMA.

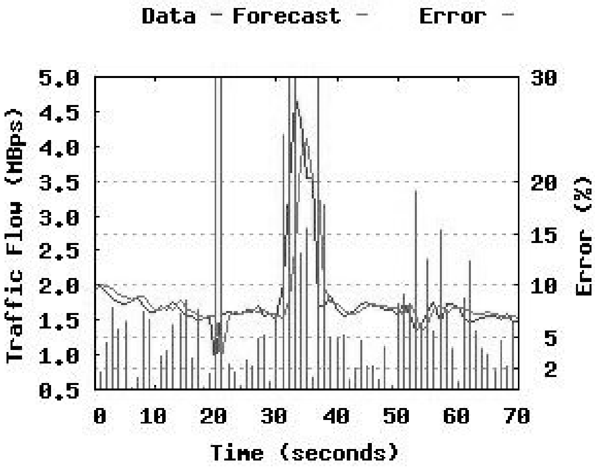


Figure 5. Output during a suddent increase in traffic (simulated DOS attack).

V. Conclusion

This paper has presented that early warning of DoS attacks is possible with the novel use of ARIMA. The ARIMA is used to predict the expected normal traffic pattern and this is used to compare with actual traffic. The computed results clearly showed that abnormal activities can be detected by the proposed method. However, this preliminary work must be extended to further investigate issues pertaining to false alarms.

A long term study on its effectiveness in a live data center environment is planned and studies on detecting other types of attacks would be of interest to us. Other considerations would be the use of different models such as SARIMA [6] or FARIMA [8] as network traffic may be seasonal in nature.

A known limitation of this technique is that it will not augur well in a small low volume network environment with less than 1 Megabytes per second traffic flow. It is most suitable for a high traffic volume data center environment.

Acknowledgment

We would like to acknowledge Multimedia Development Corporation Sdn Bhd (MDeC), which provided the MSC Malaysia Research & Development Grant Scheme (MGS), for sponsoring this research.

References

[1] H. Debar, M. Dacier, and A. Wespi, “Towards a Taxonomy of Instrusion Detection Systems”. Computer Networks, vol. 31, pp. 805-822, Aug 1999.

[2] K. Jackson, “Intrusion Detection Systems (IDS): Product Survey,” Los Alamos National Library, LA-UR-99-3883, 1999.

[3] M. David, V. Geoffrey and S. Stefan, “Inferring Internet Denial-of-Service Activity,” Proceedings of the 10th USENIX Security Symposium, 2001.

[4] IMSL C Numerical library. http://www.vni.com/products/imsl/

[5] Box, George E. P. and and Gwilym M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, Oakland.

[6] Brockwell, Peter J. and Davis, Richard A., Introduction to Time Series and Forecasting (2nd Edition), Springer-Verlag, pp. 179-208, 2002.

[7] E.D. McKenzie , General exponential smoothing and the equivalent ARMA process. *J. Forecasting* (1984), pp. 333–344.

[8] Yantai Shu, Zhigang Jin, Lianfang Zhang, Lei Wang and Yang, O.W.W., Traffic prediction using FARIMA models, *Communications*, (1999), pp. 891-895.