PhD Forum: A System Identification Approach to Monitoring Network Traffic Security

Quentin Mayo
University of North Texas
1155 Union Circle,
Denton, Texas, 76203
Email: quentinmayo@my.unt.edu

Dr. Renee Bryce
University of North Texas
1155 Union Circle,
Denton, Texas, 76203
Email: renee.bryce@unt.edu

Dr. Ram Dantu
University of North Texas
1155 Union Circle,
Denton, Texas, 76203
Email: rdantu@unt.edu

***Abstract*—Network security is a growing area of interest for cyber systems, especially given the increasing number of attacks on companies each year. Though there are a vast amount of tools already available, System Identification (SI) complements intrusion detection systems to help manage network traffic stability. SI is the science of building mathematical models of dynamic systems. This paper introduces the use of SI for modeling network traffic and utilizes a linear time invariant model to analyze performance of real connections and attack instances. We generated several ARX models where each represented a different threat state in the network. We utilized the KDD CUP 1999s DARPA dataset to analyze the performance when dealing with different attacks. Results show that the average model fit was 84.14% when determining if the system was experiencing normal traffic. This value is promising because it shows how well the system is able to determine a network state in a given time when fed input.**

***Keywords*-System Identification, ARX, Models, Security**

I. Introduction

Cyber-security is increasingly important to companies because they suffer from frequent cyber-security attacks [1]. There are many reasons that motivate cyber-attacks. The dynamics of software and hardware change is challenging to track. An automatic analysis is necessary in learning from past configurations. Additionally, given that the actual network structure is undefined, it is important to have a set of mathematical models [2], [3], [4], [5], [6]. These and similar situations favor themselves to System Identification since dynamic behavior can be modeled. When provided with enough information of the corresponding input and output with respect to a time event-driven process, it is possible to estimate a model of a system for the purpose for control, prediction, and signal processing, simulation, or error detection [3]. If we can effectively define parameters corresponding to a system's input and capture relevant output data to express the system's characteristics, a mathematical model can be built which, can capture information on a system such as its stability and performance.

II. Background and Current Area of Research

Analyzing the dynamics of a system is an important concept because basic laws of physics can be used to express that system as a set of differential equations. This differential equation contains the output variables that are attributes of interest, which can completely characterize a system. Rather than designing the systems manually, a mathematical estimation modeling based approach can be taken. This is often referred to as System Identification(SI). Though an analysis of computer networks can be more abstract than mechanical systems, parallels between physics and network characteristics can be created. An example could be shown in the network traffic flow and fluid flow systems.

To the best of our knowledge, the use of SI has yet to be explored for intrusion and anomaly detection. However, there has been research on intrusion detection using different concpets. Kaur et al. compared different wavelet intrusion detection system tools [7]. The tools reviewed used bi-directional traffic and were successful at detecting DoS and many other flood-based attacks. The detection abilities vary between attacks. Long et al. added to a weight to determine when an intrusion occurred [8]. The major difference between their research and ours is that we are able to capture time characteristics in the dataset.

We define different characteristics when a network is in a normal state by analyzing network traffic. Likewise, we look at the number of specific types of packets to determine when something malicious such as an unauthorized probe or DOS attack occurs. We can assume that the current state of a network is a unit interval, which is a closed interval between 0 and 1. This unit interval represents the output of the system. This assumption allows us to notice at what point a system goes out of a state. With any attack, we may gather traffic parameters to determine a different ARX model. Determining the dynamics of a network is expensive and time consuming. In many cases, there might be never-before-seen anomalies on the network infrastructure. Thus, we use a statistical approach for determining features of the network. Even though many attack components exist across multiple layers, many features might be independent or dependent on each other. We use an ad-hoc approach that analyzes the dynamics of different states and features. We then use the chi-squared test to measure the dependence

978-1-4673-9300-3/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCloud.2015.85

between stochastic values involving the input and output parameters.

III. ARX Model

Many processes are linear and can be modeled well with linear models. Common models include Auto-Regressive with Exogenous Input (ARX), Auto-Regressive Moving Average with Exogenous Input (ARMAX), Box-Jenkins (BJ) model, and Output-Error Models [9]. Though there are many linear models, the ARX model has been shown to be an effective model structure with great performance [10]. The performance of a system can often be described using a black-box model such as the ARX model. The basic form of the ARX model structure is show below:

$$y(t) + A_1 y(t1) + A_2 y(t2) + ... + A_{na} y(tna) = B_0 u(t) + B_1 u(t1) + ... + B_{nb} u(tnb) + e(t) \quad (1)$$

This paper considered a multivariable input and single-output ARX model rather than a single input because in many cases, the output signal might be heavily dependent on several independent or dependent input variables. We represent the output signals as different states for each model estimated. The *y(t)* denotes the output at time *t*. *na* is the number of poles, while *nb* is the number of zeros plus 1 involved with the calculations. *nk* is the dead time in the system. It reflects the number of input that will evenly effect the output. *t-1* is the notation of the output that is previously dependent on time. *e* reflects a white disturbance value or error in the model or data set. *na* and *nb* defines the order of the ARX model while *nk* is the delay in the system. In the multiple input and single-out models, *nb* and *nk* are row vectors.The ARX model does not allow for the noise and dynamics to be modeled independently. Thus, it is important to have a low signal to noise ratio.

IV. Experiment

We utilized the KDD CUP dataset with 41 features [11]. The data is loosely event-driven, which makes it applicable to time-series SI-based analysis. We focused on that dataset because it has well defined states in the system. In the results section, we use the first ten features based on the chi square test. This process was a part of feature selection, which is a form of dimensionality reduction in which we reduce the number of random variables. Likewise, many times features in a given current attack may be loosely correlated. To focus on the benefits of SI on determining the different states in the network, we first created a random initial time start up to a given instance. We used a select features instance set between 100000 to 150000 features for both the model and validation dataset. We built the model on the first half of the event dataset and then validated it on the remaining dataset. This was because we started at a random initial time, it was adequate to see how well the model performed when the system went from a known to different unknown states.

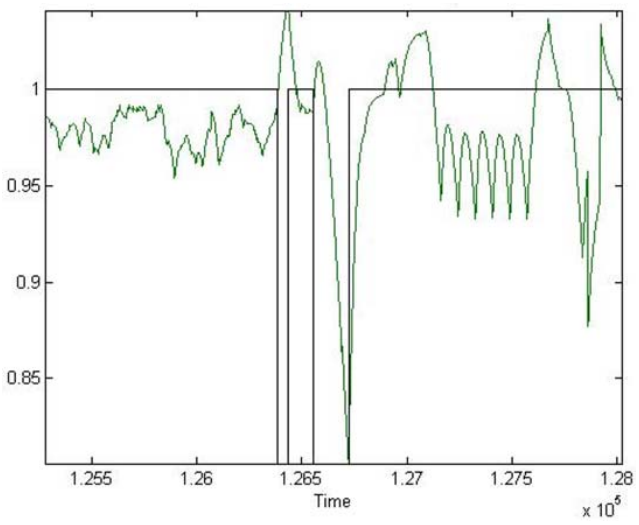

Figure 1. Subgraph of the ARX Model for Normal Traffic States

V. Results and Conclusion

In the sampled dataset, we had 6 different states corresponding to different attacks occurring on the network. One of the most important aspects of this research project is to determine when a system moves between a normal and abnormal state of operation. The best ARX model had a fit of 84.14%, but it is important to note that the model represents when a system goes in and out of a normal operation. Figure1's green line shows the change in model prediction. The model fit is based on the fit between the model prediction and the actual value using mean error squared criterion. The states move between 0 and 1, where 1 corresponds to the state that is active and 0 corresponds to the state that is not active.

This paper examines ARX models for modeling network analogies traffic but also using states to determine current system configurations. Given that we attempted to determine if the current state of the system experiences normal traffic, our model prediction was around 84.14%. The results of the model performance on the validation shows that we could visualize when an unknown state was occurring. This is the basis of anomaly detection and stability for system states. Our future work seeks to improve the performance of ARX models by exploring larger datasets over the span of longer durations. We want to improve data integrity for large datasets. We also seek to provide analysis of different attack categories. Our focus was on dynamic modeling of a network system, but we want to see how well system identification can be used for on-line learning and temporal inference. The dataset we used was loosely event driven, but more analysis should be done with using time-series datasets. The dimensionality and features selection problem should be explored more to determine the effects of increasing or

decreasing the feature space.

ACKNOWLEDGMENT

This work is supported by NSF grant (1241768).

REFERENCES

[1] P. Passeri, "Cyber attacks statistics.," *Hackmageddon. com http://hackmageddon. com/2012-cyber-attacks-statistics-master-index*, 2012.

[2] L. Andersson, U. Jönsson, K. H. Johansson, and J. Bengtsson, "A manual for system identification," *Laboratory Exercises in System Identification. KF Sigma i Lund AB. Department of Automatic Control, Lund Institute of Technology, Box*, vol. 118, 1998.

[3] K. J. Åström and P. Eykhoff, "System identificationa survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[4] B. Huang, Y. Qi, and A. Murshed, "System identification ii," *Dynamic Modelling and Predictive Control in Solid Oxide Fuel Cells: First Principle and Data-Based Approaches*, pp. 57–102, 2013.

[5] Z. Hou, Q. Shen, and H. Li, "Nonlinear system identification based on anfis," in *Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on*, vol. 1. IEEE, 2003, pp. 510–512.

[6] J.-N. Juang, "Applied system identification," 1994.

[7] G. Kaur, V. Saxena, and J. Gupta, "Anomaly detection in network traffic and role of wavelets," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 7. IEEE, 2010, pp. V7–46.

[8] J. Long, J.-p. Yin, E. Zhu, and W.-T. Zhao, "A novel active cost-sensitive learning method for intrusion detection," in *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 2. IEEE, 2008, pp. 1099–1104.

[9] H. C. Peitsman and L. L. Soethout, "Arx models and real-time model-based diagnosis," American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA (United States), Tech. Rep., 1997.

[10] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control engineering practice*, vol. 11, no. 7, pp. 733–764, 2003.

[11] M. Tavallaee, E. Bagheri, W. Lu, and A.-A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*, 2009.