An Information-Theoretic Combining Method for Multi-Classifier Anomaly Detection Systems

Ayesha Binte Ashfaq, Mobin Javed, Syed Ali Khayam
School of Electrical Engineering & Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad 44000, Pakistan
{ayesha.ashfaq, mobin.javed, ali.khayam}@seecs.edu.pk

Hayder Radha
Dept. of Electrical & Computer Engineering
Michigan State University (MSU)
East Lansing, MI 48823, USA
radha@egr.msu.edu

***Abstract*—Recent studies have shown that standalone anomaly classifiers used by network anomaly detectors are unable to provide acceptable accuracies in real-world deployments. To achieve higher accuracies, Network Anomaly Detection Systems (NADSs) now use multiple classifiers whose outputs are combined to formulate an aggregate anomaly score. Judicious methods of combining these classifiers' outputs are largely unexplored. In this paper, we propose a novel information-theoretic combining method which caters for the individual classifiers' accuracies in a multi-classifier NADS. We first show that existing combining schemes designed for or adapted to the problem of multi-classifier NADS combining do not provide good accuracies because they do not use individual classifiers' detection and false alarm rates in the combining process. Furthermore, we reveal that an accurate multi-classifier NADS, in addition to catering for the mean accuracy rates, must also consider the classifiers' variances during combining. Therefore, we propose a *Standard Deviation normalized Entropy of Accuracy* (SDnEA) method for classifier combining. Using 9 prominent classifiers operating on two publicly-available traffic datasets, we show that around $3\% - 10\%$ increase in detection rate and a 40% decrease in false alarm rate over existing combining techniques can be provided by the proposed information-theoretic NADS combining technique.**

I. Introduction

The seminal DARPA IDS evaluation of 1999 emphasized and catalyzed a shift in focus from signature-based intrusion detection to anomaly detection which can detect zero-day (previously-unknown) attacks [1]. After ten years of sustained research in the anomaly detection domain, contemporary Network Anomaly Detection Systems (NADSs) still fall short of achieving acceptable accuracies at different deployment points and under different types/rates of attacks [2], [3]. The root cause of this problem is that contemporary anomaly detectors–especially non-proprietary ones available in research literature[1]–are designed for specific traffic features which are classified using an algorithm customized for these features. Consequently, an anomaly classifier which is highly accurate for certain attacks and/or deployments fails miserably as benign or attack traffic conditions change [2].

It is now well-accepted that a NADS should use multiple anomaly classifiers to improve its accuracy (for assorted attacks) and scalability (at different deployment points and under varying traffic volumes). However, accurate and judicious methods that can be used to combine the outputs of multiple anomaly classifiers in a NADS have received little attention in research literature [8], [9].[2] The limited literature on ADS combining is either host-based [9] or relies on learning separate classifiers for different traffic classes [8]. Generic combining techniques that can combine outputs of *any* given set of NADSs are not well investigated.

In this paper, we first adapt and evaluate existing techniques that can generically combine multiple NADSs' anomaly scores. These generic techniques include: 1) Simple voting-based combining (single instance, all instances, majority vote) [5]; 2) two variants (sum and median rules) of the Bayesian pattern recognition combining method [10]; and 3) the ENCORE fusion logic from the character recognition domain [11], [12]. To quantify the improvements provided by these combining methods, we evaluate progressive combinations of 9 prominent NADS classifiers [14]–[22] on two publicly-available portscan attack datasets [2].

Our experimental evaluation shows that the accuracies of existing combining techniques have a significant room for improvement. We also reveal that, in addition to mean accuracies (detection and false alarm rates) of the NADSs, we must cater for the variance of each detector during NADS combining. Using mean and variance of accuracy values from the initial supervised learning phase of a NADS, we propose a novel information-theoretic NADS combining logic referred to as the *Standard Deviation normalized Entropy of Accuracy* (SDnEA) method. We show that SDnEA's performance consistently and considerably surpasses the accuracy of the best (ENCORE) existing combining technique. Specifically, SDnEA provides a $3\% - 10\%$ increase in detection rates and a 40% decrease in false alarm rates over ENCORE. We also show that an increase in the number of anomaly classifiers does not always induce a proportional increase in system accuracy. Therefore, a few judiciously selected classifiers can provide better system-level accuracy than many diverse classifiers.

The rest of the paper is organized as follows: Section II

This work is supported in part by the Pakistan National ICT R&D Fund and the Higher Education Commission (HEC) of Pakistan.

[1] Commercial NADSs without exception use multiple simultaneously-operating anomaly classifiers.

[2] Some prior research efforts [4],[5] have, however, proposed combining methods for signature detectors.

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

gives a brief overview of the datasets and NADS classifiers used in this study. Section III provides a comparative performance evaluation of existing combining schemes using the best operating point on the ROC curve of the respective classifiers. We then propose the SDnEA method in Section IV. This section also compares the accuracy of SDnEA with ENCORE. Section V summarizes key conclusions of this paper.

II. Datasets and Anomaly Detectors

In this study we use two real traffic datasets, independently collected at different deployment points: Router-based LBNL dataset and the Endpoint based WiSNet Lab dataset. In this section, we provide a brief overview of the datasets used in this paper; more details can be found in [2].

A. LBNL Dataset

This portscan attack dataset [7] was obtained from two international network locations at the Lawrence Berkeley National Laboratory (LBNL), USA on three distinct days. Attack traffic was isolated by identifying scans in the aggregate traffic traces. Scans were identified by flagging those hosts which unsuccessfully probed more than 20 hosts, out of which 16 hosts were probed in ascending or descending order [23]. The attack rate is significantly lower than the background traffic rate. Thus these attacks can be considered low rate relative to the background traffic rate. Moreover, a large variance can be observed in the background traffic rate at different dates.

B. Endpoint Dataset

This dataset comprises 14 months of traffic traces on a diverse set of 13 endpoints. The users of these endpoints included home users, university (researchers/students) and office users. For this study, we use 6 weeks of endpoint traffic data [6] for training and testing.

To generate attack traffic, we infected VMs on the endpoints by the following malware: `Zotob.G`, `Forbot-FU`, `Sdbot-AFR`, `Dloader-NY`, `SoBig.E@mm`, `MyDoom.A@mm`, `Blaster`, `Rbot-AQJ`, and `RBOT.CCC`. These malware have diverse scanning rates and attack ports/applications. For completeness, we also simulated three additional worms that are somewhat different from the ones described above, namely `Witty`, `CodeRedv2` and a fictitious TCP worm with a fixed and uncommon source port.

C. Anomaly Classifiers

For accuracy evaluation, we progressively combined the following prominent and diverse network anomaly classifers: 1) Rate Limiting, 2) Threshold Random Walk (TRW), 3) Credit-based Threshold Random Walk (TRW-CB), 4) Packet Header Anomaly Detector (PHAD), 5) Network Traffic Anomaly Detector (NETAD), 6) Subspace Method, 7) Kalman Filter, 8) Maximum Entropy Detector, and 9) Next-Generation Intrusion Detector (NIDES). We do not provide details of these classifiers due to brevity constraints; interested readers are referred to [14]–[22] for details.

We comprehensively compare the accuracies of different combining methods using Receiver Operating Curves (ROCs). Henceforth, all results are reported for the best operating points on these ROC curves.

III. Performance Evaluation of Existing Combining Techniques

In this section, we empirically evaluate the accuracies of existing combining methods. To maintain a logical flow of thought, description and adaptation of existing combiners to the present traffic anomaly detection problem are provided inline. In addition to gauging the accuracies of existing combining methods, we expect the results of this section to reveal insights that can be used to develop an accurate combiner.

A. Existing Combining Schemes

1) Voting-based Combining: Without loss of generality, let us treat each constituent anomaly classifier as a black-box that takes the traffic z_t as input at discrete time instance t and then outputs a binary label, [1: anomaly, 0: benign]. Let the total number of classifiers to be combined is N and let $\Lambda = \{1, 2, ..., N\}$ denote a set of indices for these classifiers. Since each classifier has two possible outcomes, it can be regarded as a binary random variable X_i. Let these variables are combined into a single anomaly score using a weighted sum random variable:

$$S_N = w_1X_1 + w_2X_2 + ... + w_NX_N = \sum_{i\in\Lambda} w_iX_i. \quad (1)$$

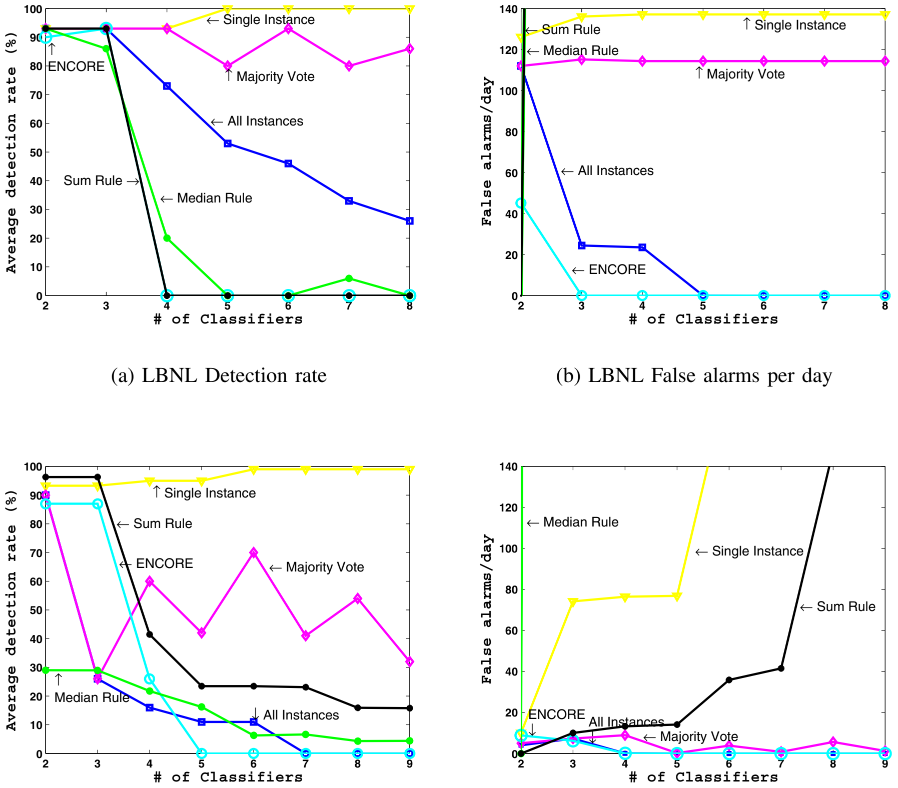
A threshold τ is then applied to this sum in order to classify a given input observation z_t as benign or anomalous.

Under a voting logic, a weight of 1 is applied to the constituent classifiers' binary scores which are then summed up: $S_t = \sum_{i\in\Lambda} X_t[i]$, where $X_t[i]$ is the output of the i-th detector for z_t. A final decision is made as follows:

$$f_{z_t}(S_t) = \begin{cases} 1 & \text{if } S_t \geq \tau_V \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

In this study we evaluate three variants of the voting principle, namely single instance ($\tau_V = 1$), majority vote ($\tau_V = N/2$), and all instance voting ($\tau_V = N$). Clearly, single instance combining will have good detection rates but poor false alarm rates. All instance voting is on the other extreme with low detection rates with low false alarm rates. Majority vote strikes a balance between these two limiting cases.

2) ENCORE Combining [12], [13]*:* ENCORE implements a decision consensus approach based on the known accuracies of the classifiers that flag an observation. In the present context, assume that after analyzing the input z_t, k out of the total N anomaly classifiers flag z_t as anomalous. ENCORE makes its combining decision based on the accuracies of the top 2 of these k classifiers. Let $p_{\alpha_t[i]}$ and $p_{\alpha_t[j]}$ respectively denote the accuracies of the top 2 classifiers, where $\alpha_t \in \{d : \text{detection}, f : \text{false alarm}\}$ can either represent detection rates or false alarm rates. The combined anomaly score is decided according to the following rule:



(a) LBNL Detection rate

(b) LBNL False alarms per day

(c) Endpoint Detection rate

(d) Endpoint False alarms per day

Fig. 1. Accuracies of existing combining methods on the LBNL and Endpoint datasets.

$$f_{z_t}(p_{\alpha_t[i]}, p_{\alpha_t[j]}) = \begin{cases} 1 & \text{if } p_{\alpha_t[i]} - p_{\alpha_t[j]} \leq \tau_E \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In other words, even an anomalous observation is classified as benign if the most accurate classifiers' accuracies are inconsistent. Note that, unlike voting-based combining, ENCORE takes the accuracies of the individual classifiers into account during the combining process. According to [11], Enhanced Majority Vote ENCORE provides the best accuracy dividends for IDS combining; details can be found in [5], [11], [12].

3) Bayesian Network based Combining: The authors in [10] combine multiple IDS classifiers, where each classifier uses its own representation of the input pattern. A Bayesian decision rule is used to identify the classifier with the maximum a posteriori probability to flag an input pattern z_t as anomalous. Assuming reasonably close a priori and a posteriori probabilities, the Bayesian detector defines two different classification rules. Under the sum rule, a sum random variable A_t is computed as function of the individual classifiers' accuracies as follows:

$$A_t = \frac{1}{k} \sum_{j=1}^{k} p_{\alpha_t[j]}, \quad (4)$$

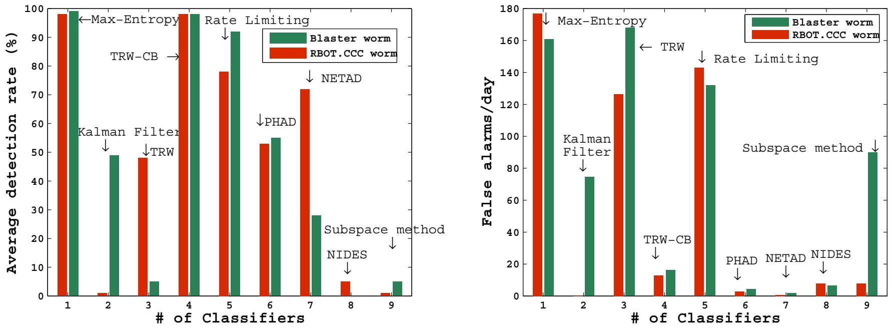
where k is the number of classifiers which flag the input as anomalous and $\alpha_t \in \{d: \text{ detection}, f: \text{ false alarm}\}$. Under the median rule, M_t for an input window z_t is the median of the individual classifies' accuracies:

$$M_t = \text{median}_{\ j=1,\ldots,k}[p_{\alpha_t[j]}]. \quad (5)$$

A_t and M_t can be thresholded ($\geq \tau_d$ in case of $\alpha_t = d$ and $\leq \tau_f$ for $\alpha_t = f$) to obtain a combined anomaly score. Similar to ENCORE, the Bayesian combiner accounts for individual classifiers' accuracies in the combining process.

B. Experimental Results

The experimental results for the voting principle, ENCORE and bayesian network based median and sum rules are shown in Fig. 1. We connect the classifiers in order of the highest performance operating points on the ROC curves. As expected, for single instance combining, as more classifiers are connected to the system, the system accuracy increases in terms of the detection rates but the false alarms also increase proportionally. The abnormal fluctuations in the behavior of the majority voting scheme is due to the non-identical nature of the classifiers that are progressively connected to the system. For the all instance voting principle, the detection rate of the system decreases as more classifiers are connected since all of them must flag the input as malicious for the system to classify it as malicious. Although this principle has a decreasing trend in the detection rates of the system, it has the same trend in its false alarms as well. ENCORE provides acceptable detection



(a) Detection rate

(b) False alarms per day

Fig. 2. Variance in the detection and false alarm rates of the classifiers on `Blaster` and `RBOT.CCC` worms.

rates since it considers each classifier's overall accuracy when combining ADSs. However, the detection rate and the false alarms go down to 0 as more and more ADSs are connected to the ensemble. The median rule offers close to 90% detection rate on the LBNL dataset, but fails to maintain this accuracy on the Endpoint dataset. The sum rule provides high detection rates on the both the datasets. However, both the techniques suffer from very high false alarm rates (2000 to 7000) on both datasets.

C. Deductions

1) Classifier Accuracies: One drawback of the voting principle based NADS combining logic is that it combines the outputs from non-identical classifiers without regard to the accuracies of the constituent classifiers. Intuitively, a classifier with higher accuracy should be given priority over a lower accuracy classifier. Hence a judicious approach is to define the weight (w_i) in (1) of each constituent classifier in accordance with its accuracy. ENCORE, considers the mean accuracies [13] of the classifiers during NADS combining and consequently offers considerably better detection rates on both datasets.

2) Variance in Accuracies: During our performance evaluation we observed that a classifier providing very good detection rate for one attack may fail completely for another attack [2]. Similarly, some classifiers' false alarms varied significantly during different periods of benign activity. Fig. 2 (a) and (b) show the detection and the false alarm rates on two specific worms, `RBOT.CCC` and `Blaster`. From these results, it is evident that the classifiers vary significantly in not only their detection rates but also in terms of their false alarms.

Hence in order to achieve good *system-level* accuracy, a NADS combiner, in addition to catering for the mean detection and false alarm rates, should also cater for the variances in the accuracies of the individual classifiers. The following section proposes an information-theoretic combiner which caters for the mean and variances of the classifiers during the combining process.

IV. An Information-Theoretic Combining Method

In this section we propose a novel NADS combining scheme based on the insights of the preceding section. Accuracy of the proposed combiner is compared with the best performing ENCORE method.

A. Combining Model

As stated in Section III-C, the weights assigned to the outputs of each classifier should characterize the accuracy of the classifier. Thus higher (lower) weights should be assigned to more (less) accurate classifiers. To this end, we propose the use of the information-theoretic Entropy measure to define the weight of a classifier. We compute a classifier's weight using its accuracy's entropy as follows:

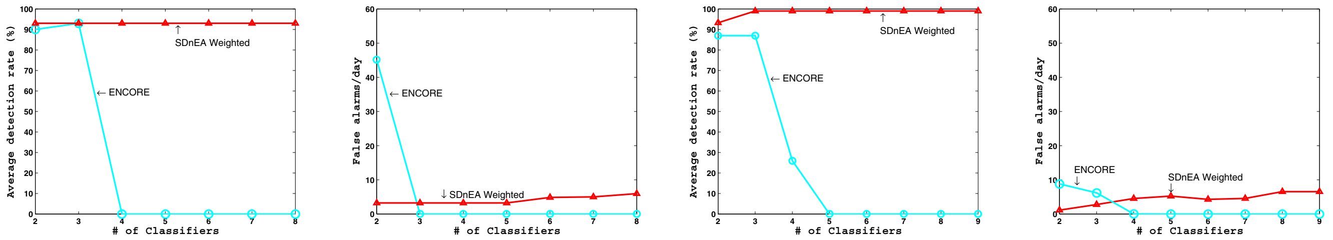
$$\begin{aligned} w_{d_i} &= 1 + p_{d_i}\log_2(p_{d_i}) + (1-p_{d_i})\log_2(1-p_{d_i}), \text{ or} \\ w_{f_i} &= -p_{f_i}\log_2(p_{f_i}) - (1-p_{f_i})\log_2(1-p_{f_i}). \end{aligned} \qquad (6)$$

Based on whether bounds on detections or false alarms are desired, one of the above weights can be used in (1) to combine multiple anomaly classifiers. Based on the above weights, higher the detection rate for a classifier, lower is the entropy value for it and higher is the weight assigned to its output. Similarly, higher are the false alarms for a classifier, lower is the entropy and a smaller weight is assigned to it.

Section III-C also revealed that connecting classifiers having low variance and acceptable detection rates can allow a system to achieve higher system-level accuracy. To incorporate classifiers' variances in the weighting process, we extend the entropy-based weighted averaging as follows:

$$\begin{aligned} w_{d_i} &= \frac{1 + p_{d_i}\log_2(p_{d_i}) + (1-p_{d_i})\log_2(1-p_{d_i})}{\sigma_{d_i}}, \text{ or} \\ w_{f_i} &= \frac{-p_{f_i}\log_2(p_{f_i}) - (1-p_{f_i})\log_2(1-p_{f_i})}{\sigma_{f_i}}, \end{aligned} \qquad (7)$$

where σ_{α_i} is the classifier's standard deviation in detection/false alarm rates. We refer to this weighting scheme as the *Standard Deviation normalized Entropy of Accuracy (SDnEA)* combining scheme.



(a) LBNL Detection rate

(b) LBNL False alarms per day

(c) Endpoint Detection rate

(d) Endpoint False alarms per day

Fig. 3. SDnEA and ENCORE accuracy comparison on the LBNL and Endpoint datasets.

B. Performance Evaluation

Fig. 3 compares the proposed SDnEA weighted scheme with ENCORE, which provided the best accuracy results in Section III. Clearly, the proposed SDnEA weighted scheme outperforms ENCORE on both LBNL and Endpoint datasets. Initially, with two to three classifier combinations, ENCORE and SDnEA provide comparable detection rates on the LBNL dataset. However, with increasing number of classifiers, SDnEA provides a 40% decrease in false alarms over ENCORE. On the endpoint dataset, SDnEA provides a 12% increase in detection rate and a 3% decrease in false alarms for the same classifier combination. These results show that, while combining multiple classifiers to achieve acceptable system performance, variance of the constituent classifiers should also be considered since it has considerable impact on the overall system accuracy.

We also note from Fig. 3 that increasing the number of classifiers in the system N does not necessarily induce a proportional increase in accuracy. Hence, increasing system complexity does not guarantee a similar increase in system accuracy. Hence, instead of relying on the number of classifiers to increase system accuracy, one should employ a few classifiers having high accuracies and low variances.

V. Conclusion

In this paper, we presented analysis of a combination of N parallel connected anomaly classifiers. We adapted and evaluated existing combining methods for the traffic anomaly detection problem and showed that the accuracies of these detectors can be improved. We then proposed a Standard Deviation normalized Entropy of Accuracy (SDnEA) combining method which provided consistent and considerable accuracy (detection rates and false alarm) improvements over existing combiners. We also showed that increasing the number of classifiers does not induce a proportional increase in system accuracy. Therefore, a few judiciously selected classifiers can provide better system-level accuracy than many diverse classifiers.

References

[1] R.P. Lippmann, J.W. Haines, D.J.Fried, J. Korba, K. Das, "The 1999 DARPA OffLine Intrusion Detection Evaluation," *Comp. Networks*, 34(2):579-595, 2000.

[2] A.B. Ashfaq, M. Joseph, A. Mumtaz, M.Q. Ali, A. Sajjad and S.A. Khayam, "A Comparative Evaluation of Anomaly Detectors under Portscan Attacks," *RAID*, 2008.

[3] C. Wong, S. Bielski, A. Studer, C. Wang, "Empirical Analysis of Rate Limiting Mechanisms," *RAID*, 2005.

[4] N. Hatami and R. Ebrahimpour, "Combining Multiple Classifiers: Diversify with Boosting and Combining by Stacking," *Intl Jrnl of Computer Science & Network Security (IJCSNS)*, (7)1, 2007.

[5] L. Xu, A. Krzyzak, C.Y. Suen, "Methods of combining multiple classifiers and their applications to Handwriting Recognition," *IEEE Trans on System, Man and Cybernetics*, (22)3, 1992.

[6] WiSNet LAB Dataset: http://www.wisnet.seecs.nust.edu.pk/downloads.php

[7] LBNL/ICSI Enterprise Tracing Project, http://www.icir.org/enterprise-tracing/download. html

[8] G. Giacinto, R. Perdisci , M. Del Rio and F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, *Information Fusion*, (9)1:69-82, 2008.

[9] S.L. Scott, "A Bayesian paradigm for designing intrusion detection systems," *Comput Stat and Data Analysis*, 45(1):6983, 2004.

[10] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, "On combining classifiers," *IEEE Trans on Pattern Analysis and Machine Intelligence*, 20(3):226239, 1998.

[11] A.F.R. Rahman, H. Alam, M.C. Fairhurst, "Multiple Classifier Combination for Character Recognition: Revisiting the Majority Voting System and Its Variations," *Intl Workshop on Doc Analysis Sys*, 2002.

[12] M. C. Fairhurst and A. F. R. Rahman, "Enhancing Consensus in Multiple Expert Decision Fusion," *IEE Proc. on Vision, Image and Signal Proc*, 147(1):3946, 2000.

[13] A. F. R. Rahman and M. C. Fairhurst, "Exploiting second order information to design a novel multiple expert decision combination platform for pattern classification," *Electronics Letters*, 33(6):476477, 1997.

[14] M.M. Williamson, "Throttling viruses: Restricting propagation to defeat malicious mobile code," *ACSAC*, 2002.

[15] J. Jung, V. Paxson, A.W. Berger, H. Balakrishnan, "Fast portscan detection using sequential hypothesis testing," *IEEE Symp Sec and Priv*, 2004.

[16] S.E. Schechter, J. Jung, A.W. Berger, "Fast detection of scanning worm infections," *RAID*, 2004.

[17] M.V. Mahoney, P.K. Chan, "PHAD: Packet Header Anomaly Detection for Indentifying Hostile Network Traffic," *Florida Tech. Technical Report*, CS-2001-4, 2001.

[18] M.V. Mahoney, P.K. Chan, "Network Traffic Anomaly Detection Based on Packet Bytes," *ACM SAC*, 2003.

[19] A. Lakhina, M. Crovella, C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM*, 2004

[20] A. Soule, K. Salamatian, N. Taft, "Combining Filtering and Statistical methods for anomaly detection," *ACM/Usenix IMC*, 2005.

[21] Y. Gu, A. McCullum, D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," *ACM/Usenix IMC*, 2005.

[22] Next-Generation Intrusion Detection Expert System (NIDES), Available: http://www.csl.sri.com/projects/nides/.

[23] R. Pang, M. Allman, V. Paxson, J. Lee, "The Devil and Packet Trace Anonymization," *ACM CCR*, 36(1), 2006.