Paper

Anomaly Detection Framework Based on Matching Pursuit for Network Security Enhancement

Rafał Renk and Witold Hołubowicz

ITTI Ltd., Poznań, Poland
Adam Mickiewicz University, Poznań, Poland

Abstract—In this paper, a framework for recognizing network traffic in order to detect anomalies is proposed. We propose to combine and correlate parameters from different layers in order to detect 0-day attacks and reduce false positives. Moreover, we propose to combine statistical and signal-based features. The major contribution of this paper are: novel framework for network security based on the correlation approach as well as new signal based algorithm for intrusion detection using matching pursuit.

Keywords—anomaly detection, intrusion detection, matching pursuit, network security, signal processing.

1. Introduction and Motivation

Intrusion detection systems (IDS) are based on mathematical models, algorithms and architectural solutions proposed for correctly detecting inappropriate, incorrect or anomalous activity within a networked systems. Intrusion detection systems can be classified as belonging to two main groups depending on the detection technique employed: anomaly detection and signature-based detection. Anomaly detection techniques, that we focus on in our work, rely on the existence of a reliable characterization of what is normal and what is not, in a particular networking scenario. More precisely, anomaly detection techniques base their evaluations on a model of what is normal, and classify as anomalous all the events that fall outside such a model. If an anomalous behavior is recognized, this does not necessarily imply that an attack activity has occurred: only few anomalies can be actually classified as attempts to compromise the security of the system.

Anomaly detection systems can be classified according to:

- the used algorithm,
- analyzed features of each packet singularly or of the whole connection,
- the kind of analyzed data - whether they focus on the packet headers or on the payload.

Most current IDS systems have problems in recognizing new attacks (0-day exploits) since they are based on the signature-based approach. In such mode, when system does not have an attack signature in database, such attack is not recognized. Another drawback of current IDS systems is that the used parameters and features do not contain all the necessary information about traffic and events in the network.

Therefore, in this paper we present the framework in which anomaly detection system based on correlation and diversity approaches are used, such as:

- Item diversity – different network layers parameters are monitored and used. In such approach we do not have information from transport layer only – such information is merged/correlated with application layer events.
- Correlation – correlation is used twofold (during decision):
  - item both anomaly and signature-based approaches are correlated,
  - parameters/features from various network layers are correlated,
  - statistical and signal-based features are used and correlated.

2. Technical Solution

In this paper, a new solution for aanomaly detection system (ADS) based on signal processing algorithm is presented. ADS analyzes traffic from Internet connection in certain point of a computer network. The proposed ADS system uses redundant signal decomposition method based on matching pursuit algorithm. ADS based on matching pursuit uses dictionary of base functions (BFD) to decompose input 1D traffic signal (1D signal may represent packets per second) into set of based functions called also atoms. The proposed BFD has a ability to approximate traffic signal. Number and parameters of base functions was limited in order to shorten atom search time process.

Since some attacks are visible only in specific layer (e.g., SQLIA), in our approach, we propose to use network parameters from different layers.

Transport layer, network layer and application layer parameters are used.

In the further step, we use the presented parameters to calculate characteristics (features) of the observed traffic. Some of the parameters are used for statistical features calculation and/or for signal-based feature calculation respectively. Feature extraction methods are presented in the following subsections.

2.1. Statistical Features

The chi-square multivariate test for anomaly detection systems can be represented by:

$$X^2 = \sum_{i=1}^{p} \frac{(X_i - \overline{X}_i)^2}{\overline{X}_i}\,, \tag{1}$$

where $X = (X_1, X_2, \ldots, X_p)$ denote an observation of p variables from a process at time t and $\overline{X} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)$ is the sample mean vector.

Using only the mean vector in Eq. (1), cause that chi-square multivariate test detects only the mean shift on one or more of the variables.

2.2. Signal Processing Features

Signal processing techniques have found application in network intrusion detection systems because of their ability to detect novel intrusions and attacks, which cannot be achieved by signature-based approaches [1]. It has been shown that network traffic presents several relevant statistical properties when analyzed at different levels (e.g., self-similarity, long range dependence, entropy variations, etc.) [2].

Approaches based on signal processing and on statistical analysis can be powerful in decomposing the signals related to network traffic, giving the ability to distinguish between trends, noise, and actual anomalous events. Wavelet-based approaches, maximum entropy estimation, principal component analysis techniques, and spectral analysis, are examples in this regard which have been investigated in the recent years by the research community [3], [4], [5], [6], [7]. However, discrete wavelet transform provides a large amount of coefficients which not necessarily reflect required features of the network signals.

Therefore, in this paper we propose another signal processing and decomposition method for anomaly/intrusion detection in networked systems. We developed original anomaly detection type IDS algorithm based on matching pursuit.

In the rest of the paper, our original ADS method will be presented in details. Moreover, results of experimental setup will be given. We tested our method with standard traces in worm detection scenario as well as in anomaly detection scenario. Discussion on redundant dictionary parameters and final conclusions will be provided.

Matching pursuit

Matching pursuit signal decomposition was proposed by Mallat and Zhang [8].

Matching pursuit is a greedy algorithm that decomposes any signal into a linear expansion of waveforms which are taken from an over complete dictionary D. The dictionary D is an over complete set of base functions called also atoms.

$$D = \{\alpha_\gamma : \gamma \in \Gamma\}\,, \tag{2}$$

where every atom α_γ from dictionary has norm equal to 1, $\|\alpha_\gamma\| = 1$, Γ represents set of indexes for atom transformation parameters such as translation, rotation and scaling.

Signal s has various representations for dictionary D. Signal can be approximated by set of atoms α_k from dictionary and projection coefficients c_k

$$s = \sum_{n=0}^{|D|-1} c_k \alpha_k\,. \tag{3}$$

To achieve best sparse decomposition of signal s (min) we have to find vector c_k with minimal norm but sufficient for proper signal reconstruction. Matching pursuit is a greedy algorithm that iteratively approximates signal to achieve good sparse signal decomposition. Matching pursuit finds set of atoms α_{γ_k} such that projection of coefficients is maximal. At first step, residual R is equal to the entire signal $R_0 = s$.

$$R_0 = \langle \alpha_{\gamma_0}, R_0 \rangle \alpha_{\gamma_0} + R_1\,. \tag{4}$$

If we want to minimize energy of residual R_1 we have to maximize the projection. $|\langle \alpha_{\gamma_0}, R_0 \rangle|$. At next step we must apply the same procedure to R_1

$$R_1 = \langle \alpha_{\gamma_1}, R_1 \rangle \alpha_{\gamma_1} + R_2\,. \tag{5}$$

Residual of signal at step n can be written

$$R^n s = R^{n-1} s - \langle R^{n-1} s \big| \alpha_{\gamma_k} \rangle \alpha_{\gamma_k}\,. \tag{6}$$

Signal s is decomposed by set of atoms

$$s = \sum_{k=0}^{N-1} \langle \alpha_{\gamma_k} | R^n s \rangle \alpha_{\gamma_k} + R^n s\,. \tag{7}$$

Algorithm stops when residual $R^n s$ of signal is lower then acceptable limit.

Our approach to intrusion detection algorithm

In basic matching pursuit algorithm atoms are selected in every step from entire dictionary which has flat structure. In this case algorithm causes significant processor burden. In our coder dictionary with internal structure was used. Dictionary is built from:

- atoms,
- centered atoms.

Centered atoms groups such atoms from D that are as more correlated as possible to each other. To calculate measure of correlation between atoms function $o(a,b)$ can be used [9]

$$o(a,b) = \sqrt{1 - \left(\frac{|\langle a,b \rangle|}{\|a\|_2 \|b\|_2} \right)^2}\,. \tag{8}$$

The quality of centered atom can be estimated according to

$$O_{k,l} = \frac{1}{|LP_{k,l}|} \sum_{i \in LP_{k,l}} o\big(A_{c(i)}, W_{c(k,l)}\big)\,, \tag{9}$$

where: $LP_{k,l}$ is a list of atoms grouped by centered atom, $O_{k,l}$ is mean of local distances from centered atom $W_{c(k,l)}$ to the atoms $A_{c(i)}$ which are strongly correlated with $A_{c(i)}$.

Centroid $W_{c(k,l)}$ represents atoms $A_{c(i)}$ which belongs to the set $i \in LP_{k,l}$. List of atoms $L_{k,l}$ should be selected according to the equation:

$$\max_{i \in LP_{k,l}} o\big(A_{c(i)}, W_{c(k,l)}\big) \leq \min_{t \in D \setminus LP_{k,l}} o\big(A_{c(t)}, W_{c(k,l)}\big)\,. \tag{10}$$

In the proposed IDS solution 1D real Gabor base function (equation was used to build dictionary) [9], [10], [11]

$$\alpha_{u,s,\xi,\phi}(t) = c_{u,s,\xi,\phi}\,\alpha\Big(\frac{t-u}{s}\Big)\cos\big(2\pi\xi(t-u)+\phi\big)\,, \tag{11}$$

where:

$$\alpha(t) = \frac{1}{\sqrt{s}}\,e^{-\pi t^2}\,, \tag{12}$$

$c_{u,s,\xi,\phi}$ – is a normalizing constant used to achieve atom unit energy.

In order to create over complete set of 1D base functions dictionary D was built by varying subsequent atom parameters: frequency ξ and phase ϕ, position u, scale s.

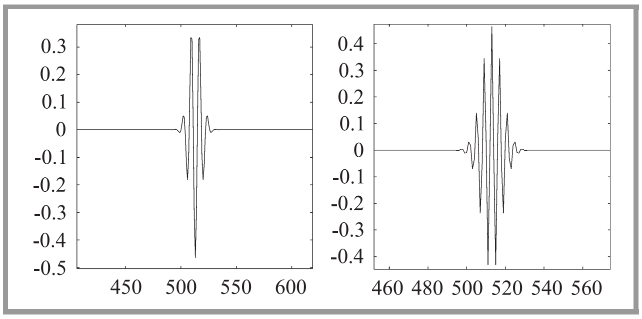

Fig. 1. Example atom from dictionary.

Base functions dictionary D was created with using 10 different scales (dyadic scales) and 50 different frequencies. In Fig. 1 example atoms from dictionary D are presented.

3. Experimental Results

Percentage of the recognized anomalies as a function of encoded atoms from dictionary of base functions is presented in Fig. 2. Five dictionaries with different parameters (different number of scales and frequencies) were used in our ADS system.

Percentage of the recognized anomalies for dictionary of base functions with approximately constant number of atoms is presented in Fig. 3. In this case we try to leave approximately constant number of atoms in dictionary but with different proportions of scales and frequencies.

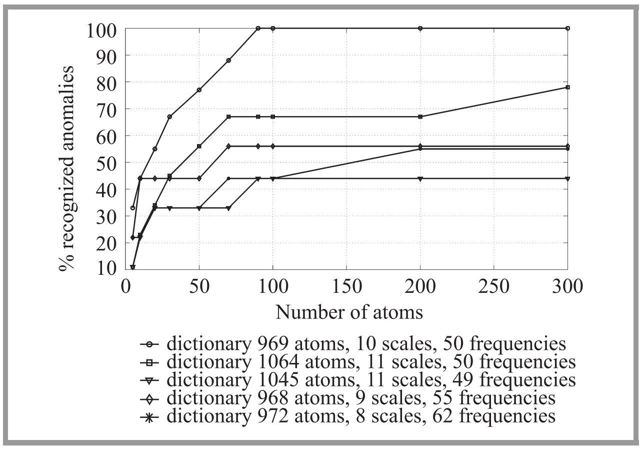

Fig. 2. Percentage of the recognized anomalies as a function of encoded atoms.

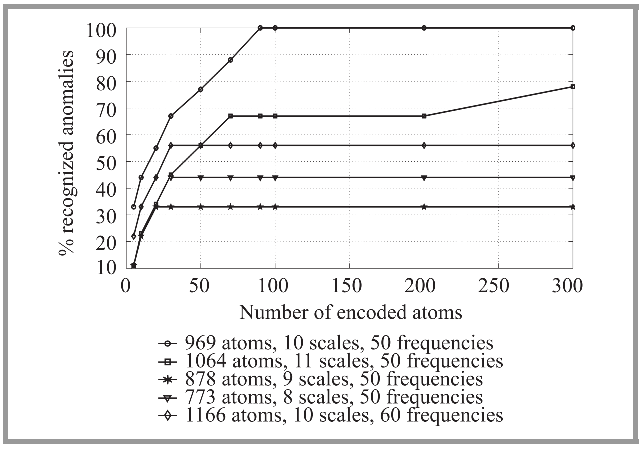

Fig. 3. Percentage of the recognized anomalies for Dictionary of Base Functions with approximately constant number of atoms.

In Tables 1, 2, 3, 4 there are example results taken from our ADS system. Traffic traces were analyzed by proposed ADS with the use of 20 minutes windows (most attacks (more than 80%) last no longer then 20 minutes). In every window we calculate matching pursuit mean projection parameter in order to recognize suspicious traffic behavior. Analyzed traces are infected by worms (Tables 1 and 2), DDos (Table 4) and DDoS SYNFlood (Table 3) attacks.

4. Conclusions

In this paper a framework for recognizing attacks and anomalies in the computer networks is presented. Our methodology is based on both statistical and signal based features. The major contribution and innovation is the application of matching pursuit algorithm to calculate network

Table 1
Matching pursuit mean projection for TCP trace [12]. Traces are analyzed with the use of 20 minutes windows

TCP trace	Window1 MP-MP	Window2 MP-MP	Window3 MP-MP	Mean MP-MP for trace	Mean MP-MP for normal trace
Mawi 2004.03.06 tcp	210,34	172,58	239,41	245,01	240,00
Mawi 2004.03.13 tcp	280,01	214,01	215,46	236,33	240,00
Mawi 20.03.2004 (attacked: worm Witty)	**322,56**	**365,24**	**351,66**	346,48	240,00
Mawi 25.03.2004 (attacked: worm Slammer)	**329,17**	**485,34**	**385,50**	400,00	240,00

Table 2
Matching pursuit mean projection for UDP trace [12]. Traces are analyzed with the use of 20 minutes windows

UDP trace	Window1 MP-MP	Window2 MP-MP	Window3 MP-MP	Mean MP-MP for trace	Mean MP-MP for normal trace
Mawi 2004.03.06 tcp	16,06	13,80	17,11	15,65	16,94
Mawi 2004.03.13 tcp	20,28	17,04	17,40	18,24	16,94
Mawi 20.03.2004 (attacked: worm Witty)	**38,12**	**75,43**	**61,78**	58,44	16,94
Mawi 25.03.2004 (attacked: worm Slammer)	**56,13**	**51,75**	**38,93**	48,93	16,94

Table 3
Matching pursuit mean projection for TCP trace [13] (traces consist of DDoS SynFlood attacks). Traces are analyzed with the use of 20 minutes windows

TCP trace	Window1 MP-MP	Window2 MP-MP	Window3 MP-MP	Mean MP-MP for trace	Mean MP-MP for normal trace
One hour trace from unina1	**1211**	**3271**	**3007**	**2496,333**	860,00
One hour trace from unina2	**1906**	**1804**	**1251**	**1653,667**	860,00

Table 4
Matching pursuit mean projection for TCP trace [14] (traces consist of DDoS attacks). Traces are analyzed with the use of 20 minutes windows

TCP trace	Window1 MP-MP	Window2 MP-MP	Window3 MP-MP	Mean MP-MP for trace	Mean MP-MP for normal trace
Backscatter 2008.11.15	147,64	**411,78**	**356,65**	305,35	153,66
Backscatter 2008.08.20	**208,40**	161,28	153,47	174,38	153,66

traffic features. The effectiveness of the proposed approach has been proved in attack and anomaly detection scenarios. Our framework can be applied to enhance military networks since it uses signal-based features. Such features can be calculated for encrypted traffic since flow characteristics are extracted without considering the payload. Future work focuses on algorithms optimization so that our framework can be applied to real-time network security enhancement and to protect federated network systems (e.g., in national project SOPAS).

Acknowledgment

Within the project supported by Polish Ministry of Science and Higher Education funds allocated for the years 2010–2012 (Research Project number OR00012511).

References

[1] M. Esposito, C. Mazzariello, F. Oliviero, S. Romano, and C. Sansone, “Real time detection of novel attacks by means of data mining techniques”, *Enterprise Information Systems*, VII 2006, Part 3, pp. 197–204.

[2] M. Esposito, C. Mazzariello, F. Oliviero, S. Romano, and C. Sansone, "Evaluating pattern recognition techniques in intrusion detection systems", in *Proc. 5th Int. Worksh. Pattern Recogn. Inf. Sys. PRIS 2005*, Miami, USA, 2005, pp. 144–153.

[3] C.-M. Cheng, H. T. Kung, , K.-S. Tan, "Use of spectral analysis in defense against DoS attacks", in *Proc. IEEE Glob. Telecommun. Conf. GLOBECOM'02*, Taipei, Taiwan, 2002, vol. 3, pp. 2143–2148.

[4] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network track anomalies", in *Proc. Internet Measur. Worksh. ACM SIGCOMM 2002*, Pittsburg, USA, 2002.

[5] P. Huang, A. Feldmann, and W. Willinger, "A non-intrusive, wavelet-based approach to detecting network performance problems", in *Proc. Internet Measur. Worksh. ACM SIGCOMM 2001*, San Diego, USA, 2001.

[6] L. Li and G. Lee, "DDoS attack detection and wavelets" in *Proc. 12th Int. Conf. Comp. Commun. Netw. ICCCN'03*, Dallas, USA, 2003, pp. 421–427.

[7] A. Dainotti, A. Pescape, and G. Ventre, "NIS04-1: wavelet-based detection of DoS attacks", in *Proc. IEEE Glob. Telecommun. Conf. GLOBECOM'06*, San Francisco, USA, 2006, pp. 1–6.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. Sig. Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] P. Jost, P. Vandergheynst, and P. Frossard, " Tree-based pursuit: algorithm and properties", *IEEE Trans. Sig. Process.*, vol. 54, no. 12, pp. 4685–4697, 2006.

[10] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation", *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[11] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps", *IEEE Trans. Sig. Process.*, vol. 49, no. 5, pp. 994–1001, 2001.

[12] "WIDE Project: MAWI Working Group Traffic Archive" [Online]. Available: http://tracer.csl.sony.co.jp/mawi/

[13] "Network Tools and Traffic Traces", Universita' degli Studi di Napoli "Federico II" [Online]. Available: http://www.grid.unina.it/Traffic/Traces/ttraces.php

[14] C. Shanon and D. Moore, "The CAIDA Dataset on the Witty Worm", March 19–24, 2004 [Online]. Available: http://www.caida.org/passive/witty

Rafał Renk – responsible for ITTI business development and management of key projects. Since 2004 professional associated with the Adam Mickiewicz University in Poznań as a researcher and lecturer in the field of software engineering and programming languages, applications in telecommunications. He also run courses covering the elements of a non-technical work in computer science. For over 10 years executes and manages number of projects addressing topic of telecommunication networks in technical and business aspect, issues of IT systems associated with the construction of new systems, systems of distance education, data warehousing, ERP, organization security and aspects of crisis management. He participated in several international projects, including these under the 5th, 6th and 7th EC Framework Program, the PASR, Phare, Leonardo da Vinci, Force Protection. He is certificated as Lead Auditor of lead information security management system according to BS 7799. He is the author or coauthor of numerous publications and presentations at national and international conferences.

e-mail: rafal.renk@itti.com.pl
ITTI Ltd.
Rubież st 46
61-612 Poznań, Poland

Division of Applied Informatics
Faculty of Physics
Adam Mickiewicz University
Umultowska st 85
61-614 Poznań, Poland

Witold Hołubowicz, Ph.D., is a graduate of the Poznań Technical University, Electrical Engineering Faculty. His professional career is split into two areas: the academic world and consulting. He was a professor in the area of telecommunications at several universities: Poznań University of Technology, Polytechnic University in New York, Franco-Polish School of New Information and Communication Technologies (EFP) in Poznań and most recently, since 2003 at the Adam Mickiewicz University in Poznań, where he is currently the Head of the Division of Applied Informatics. Throughout all of his professional career, he has been an active consultant. He participated in or coordinated more than sixty consulting, implementation and research projects on radio communications, teleinformatics and telecommunication systems, including numerous international projects. He is the president and cofounder of ITTI Ltd.. It is an independent company started in 1996, which carries out projects, both applied research and consulting, in the area of IT and telecom.

e-mail: witold.holubowicz@itti.com.pl
ITTI Ltd.
Rubież st 46
61-612 Poznań, Poland

Division of Applied Informatics
Faculty of Physics
Adam Mickiewicz University
Umultowska st 85
61-614 Poznań, Poland