Data Mining for Anomalous Network Payload Detection

S. Mrdovic
University of Sarajevo, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
sasa.mrdovic@etf.unsa.ba

***Abstract*—Network intrusion detection based on packet payload analysis is presented. Quick overview of current IDS state of the art is given. Current prevailing methods for network intrusion detection based on packet meta data, headers, will are compared with method proposed in paper. Reasoning behind packed payload analysis for intrusion detection are presented. Application of data mining methods for packet payload analysis is considered. Issues with payload analysis, like performance and false negatives and positives, will are explained.**

***Key words* – Intrusion detection, anomaly detection, network payload analysis.**

I. Introduction

Intrusion detection is part of layered defense system that well protected computer systems should have. There are different intrusion detection methods, but most of them relay on monitoring events in computer system. Idea to monitor events is almost as old as computers but term intrusion detection and its concept were introduced in 1987 [1]. Since then a number of intrusion detection systems have been proposed and tested with various success. Sherif and Dearamon give a very detailed review of papers in the area [2].

There are two main ways to classify intrusion detection systems. First classification is based on location from which data used for intrusion detection comes from. In this classification there are two main types of intrusion detection systems: host and network. Host intrusion detection systems run on a single host, collect data and detect intrusions on that host only [3]. Network intrusion detection systems passively monitor traffic on their network segment instead of monitoring single hosts [4]. Second classification is based on the way intrusions are detected. First type, in this classification, represent signature based detectors that detect known missuses. They compare data seen with patterns of known intrusions and if they find a match a flag is raised. Signature based systems have low false positive rates, which means that it seldom happens that they signal intrusion for some benign event. On the other hand they suffer from high false negative rates, meaning they fail to recognize new attacks and carefully crafted variants of old exploits. Second type are systems based on anomaly detection that create model of normal behavior in a system and detect deviations of interest that may indicate a security breach or an attempted attack. This is the original idea that started intrusion detection attempts [3] . Anomaly detectors can detect new and completely unknown attacks but sometimes some new or unusual events can be considered intrusions when they are not and that causes high false positive rates. Good review of anomaly detection intrusion detections systems can be found in [5]. For their big potential in detecting new attacks anomaly detection intrusion systems have been intensively researched but proposed solutions are more academic then practical, so that most commercial solutions are signature based [6]. Recently some hybrid ideas surfaced that suggest using deviations from normal behavior to construct signatures for future intrusion detections [7], [8].

An additional approach to intrusion detection is based on honeypot and honeynet, computer or a group of networked computers with no production function which are used only to collect data on attempts to access them. Any access to honeypot or honeynet could be considered intrusion and can help detect the source and way of attack [9].

II. Issues

There are several issues with current intrusion detection systems (IDS). Two of them are already mentioned: false negatives and false positives. Both undermine trust in IDS. False negatives bring false sense of security in which intrusions are not detected until it is to late. False positives result in that, after a while, nobody pays attention to IDS warnings. False negatives and false positives are on the opposite sides of the scale. Bringing one of them down usually brings the other one up. No current IDS seems to offer adequate levels for both. Solution might be in integrated IDS that would include both misuse and anomaly detection IDS components. This idea will be further investigated later in the paper.

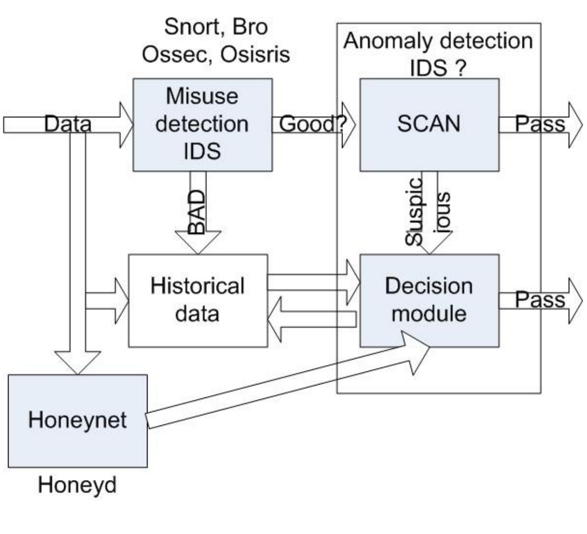
Another IDS issue is throughoutput. This means ability to analyze date needed to detect intrusion in real time without slowing down data flow or system being monitored [10]. Host IDS use host resources and their work must not have a negative impact on normal host functions. Network IDS must be able to cope with network traffic at speeds above 1 Gb/s without slowing down the traffic or missing intrusion relevant events. Most current IDS resolve this issue by analyzing only a subset of data which enables work at needed speeds but at the price of correctness of detection. Different systems use different subsets of data, but most network IDS use meta data like network packet headers [11]. Recently, an idea to analyze packet payloads have surfaced. Payload analysis would improve intrusion results, but until recently have been prohibitively slow [12]. Again integrated IDS that would include both header and payload analysis, at proper places in intrusion detection procedure, might be a solution.

III. Integrated IDS

Good sides of misuse detection IDS should be combined with good sides of anomaly detection IDS. Misuse detection IDS are excellent at detecting known attacks, and with some new ideas [13] they should be able to detect most if not all variants of those attacks. It seems obvious that data collected for intrusion detection purposes should be fed to such a signature based detector that would find and eliminate all known attacks and their variants. What is left is considered safe and should be fed to anomaly detection IDS. Traffic considered malicious by this IDS should be marked only suspicious, due to its tendency to generate false positives. Good traffic should be passed further down for production processing since nothing bad or suspicious was found and there is to best of IDS knowledge no malicious data in it. Additional help in deciding whether suspicious traffic is malicious should come from data collected from honeypot or honeynet. If suspicious traffic has been received by honeynet it means that it was not traffic for production, meaning useful and expected, purposes and should be declared bad and be dropped.

Two components of this system are available either as open source solutions or commercial products: misuse detection IDS and honeynet. The most popular open source network IDS are Snort and Bro, and most popular open source host IDS are Ossec and Osiris. Honeynets can be created using Honeyd or tools available from Honeynet project.

Component that is not readily available is anomaly detection IDS. Anomaly detection systems have been successfully implemented (at least in academic projects) in an host-based fashion, but have so far spectacularly failed to be useful in network-based systems, with a few exceptions. Specific, limited “anomaly detection” signatures have been developed and implemented in traditional commercial IDS systems, in order to detect common signs of attacks (for instance, the presence of binary codes in unusual places). However, there are no (or very few) fully fledged network-based anomaly detection systems [6]. One of the main reasons is difficulty that every (IDS) software has in detecting what is unusual and should be considered an anomaly. In its essence IDS software usually uses just some sort of statistical deviation as measure for something being unusual and potentially malicious. Model of normal behavior is usually based on historical data that is considered free of attacks which does not always have to be the case, since there might be some intrusion events that were not recognized at the time data was collected. This would lead to model that would allow future similar attacks to pass unnoticed. Also every new production service that is offered in a network would generate new and previously unseen traffic that would be different from model and would be considered suspicious. For these reasons some ideas on how these and other issues with anomaly detection IDS might be solved will be given next.



Φιγυρε1.Integrated IDS

IV. Anomaly detection process

In order to establish deviation from normal behavior it is necessary to have a model of normal behavior. Modeling of normal behavior is usually achieved by analysis of historical data on system behavior. Data mining methods are most often used for this purpose [14] . Finding events that do not fit within limits of normal is one of the task data mining is created for and various approaches in IDS area have been suggested [15], [16], [17], [18]. Other approaches used for modeling and anomaly detection include: neural networks [19], machine learning [20], Markov chain [21].

Issue of historical data not being attack free has not been addressed previously in literature. The approach this paper suggests is a simple one. Historical data should be filtered periodically. Filtration would look for intrusions in historical data and eliminate them resulting in better model of normal, intrusion free, behavior. Signature based misuse detection IDS are perfect tool for this filtration. As new intrusion signatures become available they are applied to historical data that was collected before the signature was created. This would lead to dynamical model of normal behavior that would be more accurate and more useful for representation of normal, attack free, traffic.

Another improvement in avoiding false positives for new services offered by system would be achieved by modeling those services in advance. Model of normal behavior should enable adding new traffic expected from normal service to it before it becomes available in production. This could be done by generating data for model using simulation or system other then production one that could create it. This new data should be included in normal behavior model and that would prevent most of false positives associated with new service being introduced.

V. Network packet payload analysis

As it was mentioned earlier, most current implementations in anomaly detection IDS are in academic projects and most of them are host based. There are much fewer network-based anomaly detection systems [6]. This paper tries to point some ideas for further developments in anomaly detection network IDS.

Network IDS have to deal with issue of throughoutput, they have to be able to cope with network traffic at speeds above 1 GB/s. Most of them analyze only meta data, like network packet headers, in order to be able to collect and analyze data at network speeds [11]. Packet headers provide data on which device sent data to which other device, what protocol was used, how big the packets were and similar data. In the begining network IDS this use to be enough data to create model used for anomaly detections but advances in defense mechanisms and attack paths have created the need to add another layer of defense, this one at application level that requires looking into packet payloads.

Network packet payload analysis enables detection of application level attacks. Application level attacks make a majority of current attack methods [22] for two main reasons. Most vulnerabilities that create an opportunity for attack are in user programs [23]. Most attacks at lower network communication layers, like data link, network and transport layer, can be efficiently prevented with now standard systems for network protection like firewall [24]. For those reasons advances in network packet payload analysis based intrusion detections would represent important step ahead in protection of todays computer systems.

Current trend of moving almost all applications to Web leads to need to focus the most on one of application level protocols: HTTP. This protocol is probably the most used one in current Internet oriented communications and most often one that is allowed to go through most firewalls. The fact has been used by number of attackers who used open HTTP doors to launch attacks with wider effects for them and more devastating consequences for attacked systems. Detection of HTTP attacks, especially new ones that are still not recognized by signature based IDS, seems to be the most needs computer network protection at present.

Some recent papers [25][6][13] present possible approaches to packet payload analysis. Due to amount of data that need to be analyzed it must be established if only part of the payload might be enough to decide if the packet is malicious or not. [24] finds that good results especially for HTTP payloads could be achieved using only first 185 bytes out of possible 1460. [6] suggests clustering payload before it is processed for anomaly detection. [13] approaches problem from what authors find to be the source: vulnerabilities. It tries to model vulnerability so that all the attacks based on that vulnerability can be detected.

Simple premise this paper emphasizes is that there are distinct, but not always obvious, characteristics that separate normal and malicious payloads. This assumption is just a special case of general idea on which anomaly detection IDS is based detection of behavior that differs from normal. It is also generalization of assumption that malicious behavior could be modeled too. This assumption was also used in [13] and [6]. Since intrusions are mainly, if not exclusively, based on vulnerabilities and certain types of vulnerabilities tend to repeat in various forms, idea is that model of particular types of vulnerabilities could be created that could be used to detect all intrusions based on that type of vulnerability. Good example of type of vulnerability is buffer overflow. This type of vulnerability exists from the very beginning of writing computer code, but even today the biggest percentage of vulnerabilities belong to this type [22]. All attacks based on buffer overflow vulnerabilities have characteristic group of commands that could be used to model that malicious behavior and detect future similar intrusion attempts.

Best approach to anomaly detection intrusion detection based on network packet payload analysis seems to be to use payload data filtration techniques suggested in [24] and to prepare data for data mining. Payload data should be collected and stored for off line as a historical data. Out of all payload data only particular protocol, HTTP is the best candidate, data should be extracted. Since HTTP payload is stream of ASCII characters with values between 0 and 255 each value and its position in a stream can be used as input to data mining. Each input should be declared malicious or normal using signature based intrusion detection with latest signatures as it was explained previously. In this way one all previously seen HTTP payload would be group in one of two groups.

The goal in using data mining for intrusion detection is to be able to generalize and make prediction from vast amount of historical data available. Idea is that data mining should be able to classify previously unseen input as being normal or malicious. Data mining has been used successfully in real life security do detect potential terrorist threats [26]. The key ideas are to use data mining techniques to discover consistent and useful patterns of system features that describe program and user behavior and to use the set of relevant system features to compute inductively learned classifiers that can recognize anomalies and known intrusions. Experiments have been performed using the sendmail system call data and the network tcpdump data to construct concise and accurate models to detect anomalies. Machine-learning algorithms have been used to compute the intra- and inter-audit record patterns, which are essential in describing program or user behavior. The discovered patterns can guide the audit data-gathering process and facilitate feature selection [27].

IDSs have been developed using neural networks. The idea is to train and construct a model to predict a user's next action or command based on patterns of historical behavior. The network is trained on a set of representative user commands. After the training period, the network tries to match actual commands with the actual user profile already present in the model. Any incorrectly predicted events are considered deviations from the user's established profile. Some advantages of using neural networks are that they cope well with noisy data, their success does not depend on statistical assumptions about the nature of the underlying data, and they are easier to modify for new user profiles.

IDSs have also been constructed using machine-learning algorithms to create a massive decision tree of thousands of statistical "rules" of acceptable user and system behavior. Branches on the decision tree are labeled with conditional probabilities. These machine-learning decision trees can be trained from a few days of data. However, they cannot be updated to learn new rules as usage patterns change. With these machine-learning IDSs activity is considered abnormal if it does not match a branch in the decision tree or if it matches a branch with low conditional probability.

Mentioned intrusion detection methods have not been applied to payload data in a way suggested here. Idea is to limit state space of possible input combinations to 256 by 1500 or even 200 or less. There should be plenty of input data available for data mining from historical data either collected at system being protected or from other sources like DARPA IDS test data [28]. Only issue could be amount of intrusions in historical data that might be significantly smaller than normal data. But intrusions could be artificially created in controlled environment using tools like Metasploit framework and traffic collected to provide enough data for data mining. Different lengths of packets, from only 100 to full 1460 bytes, should be tested for accurateness and speed. Result should be system that would be fed live network traffic and should be able to classify inputs, HTTP packet payloads, as being intrusion attempts or not in real time. This would provide needed anomalous payload based intrusion detection component for proposed integrated IDS.

At this point only some parts of proposed system have been tested but with promising results. To really evaluate and confirm correctness of suggested approach at least anomaly detection based on network packet payload analysis should be built. That is next step in the research being conducted. Created model will be tested with DARPA IDS test data [28] that are used for IDS testing and measurement of success. Model will be tested on real data collected from real network traffic from authors institution. Success in detecting intrusions in real time will be tested by controlled intrusion attempts based on latest attacks and by modification of known attacks. After system is tested it would become part of suggested integrated IDS for further testing of that system.

VI. Conclusion

Intrusion detection systems are evolving component of computer security. Three main issues define success of IDS: false negatives, false positives and speed. Integrated IDS that includes misuse and anomaly detection should be able to have all three of them at needed levels. Anomaly detection network IDS need most improvement. Historical data used to create model of normal behavior should be periodically filtered of intrusions as new attack signatures become available. New services that might recognized as anomalies should be modeled in advance. Network packet payload analysis would contribute the most to current intrusion detection needs. HTTP protocol data provide an avenue for attack an therefore their analysis could improve application level security significantly. Modeling of normal and anomalous HTTP payload might be achieved using data mining. Size of payload needed for successful intrusion detection might be small enough to allow real time processing.

References

[1] D. Denning, An Intrusion Detection Model, IEEE Transactions on Software Engineering, 13, 2, 222-232, 1987.

[2] J. Sherif, T. Dearmond, “Intrusion Detection: Systems and Models,” Eleventh IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'02), Pittsburgh, Pennsylvania, USA, June 10 - 12, 2002.

[3] J. P. Anderson, Computer security threat monitoring and surveillance, Technical report, J. P. Anderson Co., Ft. Washington, Pennsylvania, Apr 1980.

[4] L. Heberlein, G. Dias, K. Levitt, B. Mukherjee, J. Wood and D. Wolber, A Network Security Monitor, Proceedings of the IEEE Symposium on Research in Security and Privacy, 1990.

[5] A. Lazarevic, L. Ertoz, A. Ozgur, V. Kumar, J. Srivastava, A comparative study of anomaly detection schemes in network intrusion detection, Third SIAM International Conference on Data Mining, San Francisco, 2003.

[6] S. Zanero, S. Savaresi, Unsupervised learning techniques for an intrusion detection system, Proceedings of the 2004 ACM Symposium on Applied Computing, March, 2004.

[7] S. Rubin, S. Jha, B. P. Miller, Language-based generation and evaluation of NIDS signatures, In the IEEE Symposium on Security and Privacy, Oakland, CA, 2005

[8] J. Newsome, B. Karp, D. Song. Polygraph: Automatically Generating Signatures for Polymorphic Worms, pp. 226-241, 2005 IEEE Symposium on Security and Privacy (S&P'05), 2005.

[9] N. Provos. A Virtual Honeypot Framework. In Proceedings of the 13th USENIX Security Symposium, pages 1–14, August 2004.

[10] R.Bejtlich, The Tao of Network Security Monitoring : Beyond Intrusion Detection, Addison-Wesley Professional, July, 2004

[11] C.Kruegel, F.Valeur, G.Vigna, R.Kemmerer, Stateful Intrusion Detection for High-Speed Networks, In Proc. IEEE Symposium on Security and Privacy, 2002.

[12] C. Kruegel, T. Toth and E. Kirda, Service Specific Anomaly Detection for Network Intrusion Detection, In Symposium on Applied Computing (SAC), Spain, March 2002.

[13] D.Brumley, J.Newsome, D.Song, H.Wang, S.Jha, “Towards Automatic Generation of Vulnerability-Based Signatures”, Proceedings of the 2006 IEEE Symposium on Security and Privacy (S&P'06) - Volume 00, (pages 2 – 16), May 2006

[14] W. Lee, S. J. Stolfo, “Data Mining Approaches for Intrusion Detection”, 7th USENIX Security Symposium, San Antonio, 1998.

[15] H. S.Vaccaro and G. E. Liepins, "Detection of anomalous computer session activity,” Proceedings Symposium on Research in Security and Privacy, CA, 1989.

[16] A. Seleznyov and S. Puuronen. “Anomaly Intrusion Detection Systems: Handling Temporal Relations Between Events.” Proceedings of the Second International Workshop on Recent Advances in Intrusion Detection, W. Lafayette, IN, 1999.

[17] M. Mahoney and P. Chan. “Detecting novel attacks by identifying anomalous network packet headers.” Technical Report CS-2001-2, Florida Institute of Technology, Melbourne, FL, 2001.

[18] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. “A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data.” Technical report, CUCS, 2002.

[19] A. Ghosh and A. Schwartzbard. “A study in using neural networks for anomaly and misuse detection.” In Proceedings of the Eighth USENIX Security Symposium, 1999.

[20] C. Sinclair, L. Pierce, S.P. Matzner. “An Application of Machine Learning to Network Intrusion Detection.” 15th Annual Computer Security Applications Conference, 1999.

[21] N. Ye, “A Markov chain model of temporal behavior for anomaly detection.” Proceedings of the IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, 2000.

[22] M.Rash, A.Orebaugh, G.Clark, B.Pinkard, J.Babbin, “Intrusion Prevention and Active Response : Deploying Network and Host IPS” Syngress Publishing; 1 edition, February 1, 2005

[23] SANS Institute, “The Top 20 Most Critical Internet Security Vulnerabilities (Updated) - The Experts Consensus”, http://www.sans.org/top20/ (27.08.2006.)

[24] A.Lukatsky, "Protect Your Information With Intrusion Detection", A-List Publishing, November 2002

[25] K.Wang, S.J.Stolfo, “Anomalous payload-based network intrusion detection”, RAID Symposium, 2004

[26] K.A. Taipale, Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data, Center for Advanced Studies in Science and Technology Policy. 5 Colum. Sci. & Tech. L. Rev. 2, December 2003.

[27] Jesus Mena, Investigative Data Mining for Security and Criminal Detection, Butterworth Heinemann, 2003

[28] R. P. Lippmann et al, Evaluating Intrusion Detection Systems: The 1998 DARPA Offline Intrusion Detection Evaluation, Proceedings DARPA Information Survivability Conference and Exposition (DISCEX) 2000, Vol 2, pp. 12-26, IEEE Computer Society Press, Los Alamitos, CA, 2000.