

Combining Multiple One-Class Classifiers for Hardening Payload-based Anomaly Detection Systems (Extended Abstract)

Roberto Perdisci, Guofei Gu, Wenke Lee

College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
{rperdisc,guofei,wenke}@cc.gatech.edu

Intrusion Detection Systems (IDS) are valuable tools for the defense-in-depth of computer networks. Network IDS look for known or potential malicious activities in network traffic and raise an alarm whenever a suspicious activity is detected. Two main approaches to intrusion detection are used, namely *misuse* and *anomaly* detection [10]. Misuse detectors are based on a description of known malicious activities. This description is often modeled as a set of rules referred to as *attack signatures*. Activities that match an attack signature are classified as malicious. Anomaly detectors are based on a description of *normal* or *benign* activities. A distance between the description of normal events and new network activities is measured. As malicious activities are expected to be different from normal activities, a suitable distance measure allows anomaly-based IDS to detect attack traffic. Anomaly-based detection systems usually produce a relatively higher number of false positives, compared to the misuse-based or *signature-based* detection systems. However, anomaly detectors are able to detect *zero-day* (i.e., never-before-seen) attacks, whereas signature-based systems are not.

Unsupervised or *unlabeled* learning approaches for network anomaly detection have been proposed in [12, 4]. These methods aim to work on datasets of traffic extracted from real networks without the necessity of a labeling process. *Unlabeled* anomaly detection systems are based on the reasonable assumption that the percentage of attack patterns in the extracted traffic traces is usually much lower than the percentage of normal patterns [12]. Furthermore, it is possible to use signature-based IDS in order to filter the extracted traffic by removing the known attacks, thus further reducing the number of attack patterns possibly present in the dataset. Another assumption is that the attack patterns are supposed to be distinguishable from the normal patterns in a suitable feature space. The term “unlabeled anomaly detection” used in the intrusion detection field actually refers to what in machine learning is more often called “novelty detection”, “outlier detection” or “one-class classification”.

Recent work on unlabeled anomaly detection focused on *high speed* classification based on simple *payload** statistics [7, 9, 14, 15]. For example, PAYL [14, 15] extracts 256 features from the payload. Each feature represents the occurrence frequency in the payload of one of the 256 possible byte values. A simple model of normal traffic is then constructed by computing the average and standard deviation of each feature. A payload is considered anomalous if a *simplified Mahalanobis distance* between the payload under test and the model of normal traffic exceeds a predetermined threshold. Wang et al. [14] also proposed a more generic n -gram[†] version of PAYL. In this case the payload is described by a pattern vector in a 256^n -dimensional feature space. The n -grams extract byte sequence information from the payload, which helps in constructing a more precise model of the normal traffic compared to the simple byte frequency-based model. The extraction of n -gram statistics from the payload can be performed efficiently and the IDS can be used to monitor high speed links in real time. However, given the exponentially growing number of extracted features, the higher n the more difficult it may be to construct an accurate model because of the curse of dimensionality and possible computational complexity problems.

It has been demonstrated that many anomaly detection systems can be “evaded” by *mimicry* attacks [13, 6, 1, 5]. A mimicry attack is an attack against a network or system vulnerability that is carefully crafted so that the attack

*The *payload* is the data portion of a network packet.

†Here an n -gram represents n consecutive bytes in the payload

pattern, i.e., the representation of the attack used during the classification process, lies inside the decision surface that separates the normal patterns from the anomalous ones (i.e., the *outliers*). A successful mimicry attack is able to exploit the targeted vulnerability while causing the anomaly IDS to produce a false negative (i.e., no alarm is raised). In [5], Fogla et al. showed how to construct a mimicry attack, called *polymorphic blending attack*, that can evade 1-gram (i.e., the *single-byte frequency* version) and 2-gram PAYL. Using byte substitution and padding techniques, the polymorphic blending attack encodes the attack payload so that the obtained *transformed* attack is classified as normal by PAYL, while still being able to exploit the targeted vulnerability.

We propose a new approach to construct a *high speed* payload-based anomaly IDS by combining multiple one-class SVM classifiers. Our approach is intended to improve both the detection accuracy and the hardness of evasion of high speed anomaly detectors.

MCS attain accuracy improvements when the combined classifiers are “diverse”, i.e., they make different errors on new patterns [3]. A way to induce diversity is to combine classifiers that are based on descriptions of the patterns in different feature spaces [8]. We propose a new technique to extract the features from the payload that is similar to the 2-gram technique. Instead of measuring the frequency of the pairs of consecutive bytes, we propose to measure the features by using a sliding window that “covers” two bytes which are ν positions apart from each other in the payload. We refer to this pairs of bytes as 2_ν -grams. This feature extraction process does not add any complexity with respect to the traditional 2-gram technique and can be performed efficiently. We also show that the proposed technique allows us to “summarize” the occurrence frequency of n -grams, with $n > 2$, thus capturing byte sequence information while limiting the dimensionality of the feature space. By varying the parameter ν , we construct a representation of the payload in different feature spaces. Then we use a feature clustering algorithm originally proposed in [2] for text classification problems to reduce the dimensionality of the different feature spaces where the payload is represented. Detection accuracy and hardness of evasion are obtained by constructing our anomaly-based IDS using a combination of multiple one-class SVM classifiers that work on these different feature spaces. Using multiple classifiers forces the attacker to devise a mimicry attack that evades multiple models of normal traffic at the same time, which is intuitively harder than evading just one model.

We compared our payload-based anomaly IDS to the original implementation of 1-gram PAYL [14] by Columbia University, to an implementation of 2-gram PAYL, and to an IDS constructed by combining multiple one-class classifiers based on the *simplified Mahalanobis distance* used by PAYL. We performed our tests on 5 days of (assumed) benign HTTP traffic collected from our academic network, and on a dataset of 11 “standard” attacks, 6 polymorphic attacks generated using CLET [1], and a Polymorphic Blending Attack [5]. The complete experimental results were published in [11].

DFP(%)	RFP(%)	Detected attacks	DR(%)
0.0	0.00022	1	0.8
0.01	0.01451	4	17.5
0.1	0.15275	17	69.1
1.0	0.92694	17	72.2
2.0	1.86263	17	72.2
5.0	5.69681	18	73.8
10.0	11.05049	18	78.6

Table 1: Performance of 1-gram PAYL.

DFP(%)	RFP(%)	Detected attacks	DR(%)
0.0	0.0	0	0
0.01	0.00381	17	68.5
0.1	0.07460	17	79.0
1.0	0.49102	18	99.2
2.0	1.14952	18	99.2
5.0	3.47902	18	99.2
10.0	7.50843	18	100

Table 2: Performance of an ensemble of one-class SVM using 40 feature clusters.

Table 1 shows the results obtained using PAYL, whereas Table 2 shows the results obtained using our MCS-based IDS. Both the IDS were trained on the first day of normal traffic and tested on the next 4 days or normal traffic plus the dataset of attacks. DFP is the “desired” false positive rate, i.e., the false positive rate set as a parameter during training. The RFP is the “real” false positive rate, i.e., the false positives measured on the test dataset of normal traffic. The number of detected attacks is computed by considering an attack as detected if at least one of its attack packets is detected, whereas the detection rate (DR) is the overall percentage of detected attack packets (regardless of the attack they belong to). It is easy to see that our IDS performs better than PAYL. For a given detection rate, PAYL tends to generate more false positives than our IDS. Furthermore, we found that PAYL is not able to detect the Polymorphic Blending Attack unless we are willing to accept an RFP as high as 4.02% (below this value PAYL is able to detect only 17 out of 18 attacks), which is usually not tolerable for a network-based IDS. This confirms the effectiveness of the Polymorphic Blending Attack. On the other hand, our IDS is able to detect all the attacks, including the Polymorphic Blending Attack, with an RFP lower than 0.5%. This shows that our IDS is more robust than PAYL with respect to the Polymorphic Blending Attack.

References

- [1] T. Detristan, T. Ulenspiegel, Y. Malcom, and M. Underduk. Polymorphic shellcode engine using spectrum analysis. *Phrack Issue 0x3d*, 2003.
- [2] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.
- [3] T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS)*, 2000.
- [4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In D. Barbara and S. Jajodia, editors, *Applications of Data Mining in Computer Security*. Kluwer, 2002.
- [5] P. Fogla, M. Sharif, R. Perdisci, O. M. Kolesnikov, and W. Lee. Polymorphic blending attack. In *USENIX Security Symposium*, 2006.
- [6] C. Kruegel, E. Kirda, D. Mutz, W. Robertson, and G. Vigna. Automating mimicry attacks using static binary analysis. In *USENIX Security Symposium*, 2005.
- [7] C. Kruegel, T. Toth, and E. Kirda. Service specific anomaly detection for network intrusion detection. In *ACM Symposium on Applied Computing (SAC)*, 2002.
- [8] L. Kuncheva. *Combining Pattern Classifiers*. Wiley, 2004.
- [9] M. Mahoney. Network traffic anomaly detection based on packet bytes. In *ACM Symposium on Applied Computing (SAC)*, 2003.
- [10] J. McHugh, A. Christie, and J. Allen. Defending yourself: The role of intrusion detection systems. *IEEE Software*, pages 42–51, Sept./Oct. 2000.
- [11] Roberto Perdisci, Guofei Gu, and Wenke Lee. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 488–498, Washington, DC, USA, 2006. IEEE Computer Society.
- [12] L. Portnoy, E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In *ACM CSS Workshop on Data Mining Applied to Security*, 2001.
- [13] D. Wagner and P. Soto. Mimicry attacks on host-based intrusion detection systems. In *ACM Conference on Computer and Communication Security (ACM CCS)*, 2002.
- [14] K. Wang and S. Stolfo. Anomalous payload-based network intrusion detection. In *Recent Advances in Intrusion Detection (RAID)*, 2004.
- [15] K. Wang and S. Stolfo. Anomalous payload-based worm detection and signature generation. In *Recent Advances in Intrusion Detection (RAID)*, 2005.