Adversarial Communication Networks Modeling for Intrusion Detection Strengthened against Mimicry

Jorge Maestre Vidal
jmaestre@indra.es
Indra, Digital Lab
Alcobendas, Madrid, Spain

Marco Antonio Sotelo Monge
masotelo@ucm.es
Complutense University of Madrid
Madrid, Spain

ABSTRACT

The rapid evolution of the emerging communication landscape prompted the rise of never seen before threats, in this way encouraging the development of more effective Network-based Intrusion Detection Systems (NIDS) able to recognize outlying behaviors. But despite the theoretical effectiveness of the existing state-of-the-art, the in-depth review of the bibliography suggests the need for their constant adaptation to the changes in their operational environment and preventing being evaded by mimicry methods. The latest threats attempt to hide the malicious actions in a tangle of statistical features that simulate the normal use of the protected network, so they acquire a greater chance of avoiding the defensive actuators. In order to contribute to their mitigation, this paper introduces a novel intrusion detection strategy resistant against mimicry. The proposal constructs models of the network usage and from them, analyzes the binary contents of the traffic payload looking for outlying patterns that may evidence malicious contents. In contrast to most previous solutions, our research overcomes the traditional strengthening via randomization, by taking advantage of scoring the suspicious packet similarity between legitimate and previously built adversarial models. Its effectiveness was evaluated on the public datasets DARPA'99 and UCM 2011, where its ability to recognize attacks obfuscated by imitation was proven.

CCS CONCEPTS

• **Security and privacy** → Intrusion/anomaly detection and malware mitigation; • **Computing methodologies** → Modeling and simulation.

KEYWORDS

Adversarial Attacks, Anomalies, Communication Networks, Intrusion Detection

ACM Reference Format:
Jorge Maestre Vidal and Marco Antonio Sotelo Monge. 2019. Adversarial Communication Networks Modeling for Intrusion Detection Strengthened against Mimicry. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019) (ARES '19), Aug., Canterbury, UK*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3339252.3340335

ARES '19, Aug., Canterbury, UK

ACM ISBN 978-1-4503-7164-3/19/08…$15.00
https://doi.org/10.1145/3339252.3340335

1 INTRODUCTION

The pioneering solutions for intrusion detection based on modeling and analyzing network environments originally took advantage of pattern recognition strategies able to discover previously known evidences of attacks [7]. But the rapid proliferation of the network technologies gave rise to a massive amount of never seen before threats, in this way encouraging the development of alternative solutions able to additionally deal with novel malicious behaviors. Because of their effectiveness in this context, the anomaly-based intrusion detection paradigm has consolidated as the ground of most of the existing Network-based Intrusion Detection Systems (NIDS) [14]. This *modus operandi* typically relies on building usage models from normal/legitimate observations, then monitoring the operational environment looking for significant discordances, which are tagged as "suspicious". Among the different publications that have laid the foundations for their development, our research focused on the Anomalous Payload-Based Intrusion Detection (PAYL) method [29, 31]. PAYL focused on analyzing the traffic payload looking for statistical outliers within each packet context. But as much as it has been evolved by the research community, the in-depth review of the bibliography reveals several challenges when operating in current commutation scenarios [9, 28], like difficulties when modeling data extracted from very heterogeneous sources, high consumption of computational resources, weak adaptability to non-stationary (concept drift) and susceptibility to evasion methods based on adversarial machine learning [20], the latter being the main target of our research. In order to contribute to their mitigation, this paper introduces a novel intrusion detection strategy resistant against mimicry, built on the grounds of the PAYL sensor family, that attempts to strengthen the previously released Advanced Payload Analyzer Preprocessor (APAP) [17]. In a similar way to its predecessors, the proposal constructs models of the network usage and from them, analyzes the binary contents of the traffic payload looking for outlying patterns that may evidence malicious contents. In contrast to most previous solutions, our research overcomes the traditional strengthening via randomization, by taking advantage of scoring the suspicious packet similarity between legitimate and previously built adversarial models.

The paper is organized into five sections, being the first of them present introduction. Section 2 reviews the related works and describes the considered adversarial attack model. Section 3 presents a novel NIDS descendant of the PAYL sensor family strengthened

against evasion by adversarial tactics. Section 4 discusses the conducted experimentation and the observed results. Finally, Section 5 summarizes the conclusions and forthcoming research actions.

2 BACKGROUND

This section introduces the main features and drawbacks of the PAYL family of sensors and APAP, which are the precursors of the undertaken research.

2.1 Payload-Based Intrusion Detection

The first sensor of the PAYL family was published by Wang and Stolfo in 2004 [31]. According to their authors, it initially intended "to detect the first occurrences of a worm either at a network system gateway or within an internal network from a rogue device and to prevent its propagation". Although the problem to be solved was the worm recognition, their proposal was also valid for a wide range of intrusion attempts, thus inspiring further research. PAYL was characterized by building a light legitimate usage model from 256 interrelated features represented as 256-element histograms. They were extracted according to the N-gram [24] methodology, as 1-grams, from the payload byte frequency distribution. At detection stage, the similarity between the normal model built at training stage and the model generated from the incoming traffic was compared. If their divergence exceeds a predefined threshold, an alert was reported. In [4] some PAYL problems inherent in building models from different length payload were fixed by a novel multi-level solution driven by a Self-Organizing Map (SOM). On the other hand, their difficulties related with processing large payloads at real-time were studied in [27]. In [21] a Support Vector Machine (SVM) ensemble was proposed for enhancing the NIDS accuracy, similarly being addressed by Hidden Markov Models (HMM) in [1]. According to this study, the previous publications were not able to accurately recognize attacks like Cross Site Scripting and SQL-Injection, where the payload statistics were not significantly different from normal traffic. This problem was the main object of study of [25], where the detector effectiveness increased when implementing Multinomial Bayesian one class classification. Likewise ANAGRAM [30], these sensors took advantage of randomly sampling with performance enhancement purposes [1, 21], which became an usual strategy for reducing their Deep Packet Inspection (DPI) impact on real communication environments. This problem was also addressed by [11, 12].

2.2 Advanced Payload Analyzer Preprocessor

In response to the challenges of the emergent monitoring environments, the Advanced Payload Analyzer Preprocessor (APAP) [17] introduced a more complex variation of the ANAGRAM detector [30]. APAP was originally released as a preprocessing module of Snort, in this way combining the rule-based capabilities of the popular NIDS with a novel anomaly-based detection stage. The latest took advantage of N-gram for payload feature extraction [24]. On the other hand, their storage/access were driven by Counting Bloom Filters (CBF) [23], which reduced the host system memory consumption and enhanced their hashing. APAP comprises five different data processing stages, that are grouped into a pair of sets of actions: Training and Detection. At training, four tasks are distinguished: *Initialization*, *Base Training*, *Reference Training* and *K-values Definition*. During *Initialization*, APAP proceeds to the elimination of information from previous trainings by clearing the CBFs (al positions to 0) and establishing the proper hashing function. At *Base Training*, a CBF is filled with information extracted from normal traffic payload samples. This data structure stores the occurrence of each possible n-gram on the binary contents of the payload. At *Reference Training*, the k-values of the sensor are calculated. They are metrics that summarize the content of the CBF and facilitate the generation of detection rules, being aggregated similarly to the histogram-based approach adopted in [31]. With this purpose, a dataset of "malicious" samples is considered, which serves as filler of a duplicate of the CBF modified at Base Training. The resultant k-values are translated into detection rules at *K-values Definition*, which assumes the contrast between CBFs filled with "normal" and "malicious" contents. When APAP operates at *Detection* mode, these rules determine the triggering of alerts and/or countermeasures.

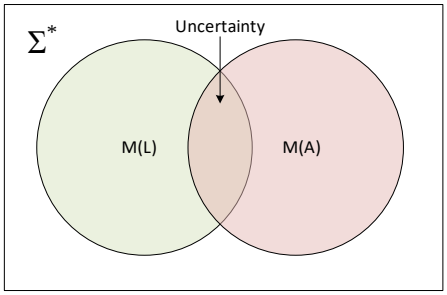

Figure 1: Observations likely to go unnoticed by NIDS.

2.3 Evasion by Mimicry

Over the past decade, the research community has varied its perception of the mimicry attacks regarding the operational environment for which they were designed. One of the preliminarily views of weaponizing mimicry for evading IDS was presented in [13]. Accordingly, this kind of threats were able to thwart IDS based on modelling and analyzing sequences of systems calls by interleaving typically legitimate actions between malicious steps. Further research like [16, 26] studied how to strengthen insider detection systems against related threats. The mimicry attacks were also in-depth reviewed in the field of recognizing intrusions by analyzing the payload contents of the network packets [20, 30], nowadays entailing an emerging challenge. Bearing in mind the mimicry attacks representation introduced in [13], these threats can be understood as actions that attempt to exploit the situation described in Fig. 1; were Σ is the set of n-grams extractable from a packet payload, and Σ^* is the infinite set of all possible payloads. A legitimate model accepts the n-grams within $M(L)$ as normal, while the n-grams within $M(A)$ are labelled as potentially harmful. Consequently, observations in the intersection between $M(L)$ and $M(A)$ will lead to non-deterministic labels, where typically, the closer to $M(L)$ the greater likelihood of being labelled as normal.

In the PAYL family context, mimicry attacks attempt to exploit the uncertainty inherent in the region $M(L) \cap M(A)$ by hiding the

malware contents within a wrapping of padding content extracted from $M(L)$. Henceforth the original attack semantic must be preserved, while the characteristics of padding must fit the normal profiles used by the detector at training stage [26]. The research towards preventing mimicry attacks widely assumed that the attacker has perfect knowledge about $M(L)$, the targeted network usage modelling and the deployed detection strategies. In addition, it assumes that attackers have the ability of crafting and poisoning malicious packets. In response to the PAYL capabilities, the attackers adopted a wide variety of ingenious and practical evasion methods [3], mimicry being one of their most outstanding tactics. As demonstrated in [6], they were able to thwart PAYL, which encouraged the design of strengthening capabilities. With this purpose, the ANAGRAM [30] proposal introduced a 2-class classifier that implemented Bloom filters [22] for registering the binary distribution of the payload, thus allowing to operate on n-grams of greater size. ANAGRAM proposed a randomized n-gram modelling, which was proved as a very accurate solution for adding difficulty to the generation of adversarial samples. In these grounds, the PAYL family has traditionally faced mimicry by randomizing the n-grams extraction process, which effectiveness heavily relied on the premise that the attack always had a small invariant part clearly distinguishable in $M(A)$. But the rapid evolution of the adversarial machine learning enablers has led to many situations where this approach becomes insufficient [20]. Bearing this in mind, the strengthening of our proposal overcomes randomization by taking advantage of scoring the suspicious packet similarity between $M(L)$ and adversarial models.

3 PAYLOAD ANALYSIS ROBUST AGAINST ADVERSARIAL THREATS

As an enhancement of the APAP/APACS methods [17, 18], the strengthening solution presented in this paper addresses the following modus operandi: at Training both normal and adversarial models are constructed according to features extracted by the N-gram methodology. At Detection stage the payloads to be analyzed are gathered from the protected environment and compared with the usage models previously built. The similarity measures between the observations and the models allow to estimate their nature (normal or suspicious) and the labeling coherency (see Fig. 2). Our research assumed that the greatest inference the higher probability of the samples being crafted by the intruders. This section describes each data processing stage and the adopted decision criteria.

3.1 Base Training

At Base Training stage, the model that summarized the main features of the legitimate payloads that flow on the defended network is built. A collection of representative samples of legitimate traffic, i.e. traces corresponding to the habitual usage of the network (and hence not discordant), is necessary for its elaboration. The normal usage model takes advantage of the CBFs for storing the frequency of occurrence of each n-gram within the payload, thus filling the $CBF(L)$ with the information extracted from the reference samples, until it is possible to conclude that its contents are representative enough. It is assumed that this occurs when new information is

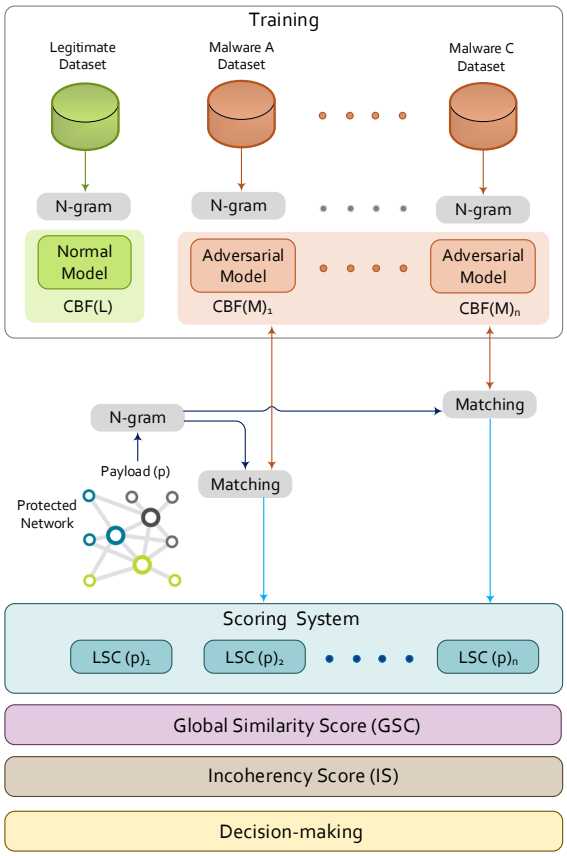

Figure 2: Data Processing Stages.

added and there are no representative variations on the data distribution within the filter. This saturation is assessed by implementing and adapting the "elbow" method [8], where knee-point is computed by observing the Sum of the Squared Errors (SSE) between positions of the $CBF(L)$ filled by normal traffic. Note that regardless the size of the n-gram window, the number of maximum steps that can be executed till reaching the aforementioned saturation point is limited by the system resources. This is avoided by directly operating with stochastic data, where each position of the filter may represent the probability of finding the n-gram within the payload of each packet of the training dataset.

3.2 Refinement by adversarial models

With the purpose of reducing the false positive rate, at the Refinement Training stage, the proposal generates adversarial traffic models able to represent the binary contents of different malware traits. By following the same procedure applied for constructing normal usage models, a CBF per group of malicious samples is properly filled. Let the n malicious traffic characterizations to be taken into account, this must result in the provisional successfully filled adversarial models:

$$CBF'(M) := CBF'(M)_1, \cdots, CBF'(M)_n \quad (1)$$

Note that the original malicious traces may contain a baseline of normal traffic on which the malware is transferred. In order to attempt to minimize the number of payload labelling errors, the subtraction operation is applied between each of them and the $CBF(L)$ normal usage model:

$$CBF(M) := CBF'(M)_1 - CBF(L), \cdots, CBF'(M)_n - CBF(L) \quad (2)$$

3.3 Scoring and Intrusion Detection

Let a traffic payload p of length m to be analyzed, $p : b_1, b_2, b_3, \cdots b_m$, each b_i, $0 < i \leq m$ represents a bit [0,1] of its contents. When assuming a k-gram sliding window for their extraction, the CBF positions indicated by the binary sequences $b_1, b_2, \cdots b_{k+1}$, etc., are accessed in $CBF(L)$ and each CBF $CBF(M)_1, \cdots, CBF(M)_n$ adversarial model. From the obtained results, the proposal calculates the packet score, which considers both binary-based and spectral-based traits. In the first case, it is obtained the number of matches (CBF positions greater than 0) regarding the normal (γ) and adversarial ($\delta_1, \ldots, \delta_n$) models. In the grounds of the spectral-based approaches, the frequencies of apparition stored in the filters are taken into account, being α regarding γ, and $\beta_1, \cdots, \beta_n$ regarding $\delta_1, \ldots, \delta_n$. Let the legitimate model $CBF(L)$ and its adversarial representation $CBF(M)_i$, $0 < i \leq n$, the Local Similarity Score (LSC) of the payload p is defined by the following expression:

$$LSC(p) = \frac{\alpha - \beta_i}{\alpha + \beta_i + \mu}, LSC \in [-1, 1] \quad (3)$$

where α is the sum of occurrences of the n-grams extracted from p matched successfully in $CBF(L)$ (when the CBF position is greater than 0), β_i is the sum of occurrences of the n-grams extracted from p matched successfully in $CBF(M)_i$, and μ is the number of n-grams with value 0 at the CBFs. On the other hand, the Global Similarity Score (GSC) of p is defined by the following expression:

$$GSC(p) = Min\{LSC_1, \cdots, LSC_n\}, GDC(p) \in [-1, 1] \quad (4)$$

At detection stage, the NIDS monitors the traffic packets payloads looking for traits of malicious contents. In particular, the NIDS calculates their Local Similarity Scores (LSC) and Global Similarity Scores (GSC) regarding the network normal usage representation $CBF(L)$ and adversarial models $CBF(M)_1, \cdots, CBF(M)_n$ previously built at *Training* stage. When $GSC(p) < \tau$ alerts are issued, being $\tau \in [-1, 1]$ a previously defined confidence interval acting as adjustment parameter of the sensor.

3.4 Strengthening against mimicry

The main advantage of the PAYL sensor family against mimicry is that at *Training* stages, it is able to separate the normal contents from the malicious samples (see Section 3.2). In our proposal, this exactly occurs once the operations $CBF'(M)_i = CBF(M)_i - CBF(L)$ are performed, where each common pattern between $CBF(L)$ and $CBF(M)_i$ was reduced or eliminated. Consequently, it is highly unexpected that a payload p to be analyzed displays great and close similitude with $CBF(L)$ and some of the $CBF(M)_1, \cdots, CBF(M)_n$ at Detection stage. We hypothesized that from this situation can be deducted that the NIDS did not label the sample with enough reliability (the facts that p seems "normal" and "malicious" at the same time are quite contradictory). Consequently, it was assumed that when the sensor operates on cleansed adversarial models, this poses a potential evidence of obfuscation by mimicry [16], or a deficient *Training* phase. Accordingly, during the experiment it was considered that an Incoherent Labelling (IL) occurred when the following first-order logical expression was satisfied:

$$\frac{\alpha}{\alpha + \mu_\alpha} > \phi \quad and \quad \exists_i, \frac{\beta_i}{\beta_i + \mu_\beta} > \phi \rightarrow IL \quad (5)$$

where α is the sum of occurrences of the n-grams extracted from p matched successfully in $CBF(L)$; (when the CBF position is greater than 0); μ_α is the number of n-grams in $CBF(L)$ as 0; β_i is the sum of occurrences of the n-grams extracted from p matched successfully in $CBF(M)_i$; and μ is the number of n-grams with value 0. The incoherency confidence interval ϕ acts as adjustment parameter of the sensor strengthening against mimicry. The Incoherency Score (IS) was calculated as follows

$$IS(p) = 1 - |\frac{\alpha}{\alpha + \mu_\alpha} - Max\{\frac{\beta_i}{\beta_i + \mu_i}\}| \quad (6)$$

4 EXPERIMENTS AND RESULTS

Thorough this sections, the datasets considered for training/evaluation are described, and the achieved results are discussed.

4.1 Datasets and evaluation methodology

The proposal was evaluated based on the DARPA'99 and UCM 2011 Datasets. DARPA'99 [15] provides online and offline collections of real/synthetic samples monitored on an experimental environment. As usual in the bibliography, the validation of proposal considered the second one [10], in this being trained based on traffic captures gathered during 7 days, separated in sessions labeled as "normal" or "attacks". Note that the "normal" ones were used for building the normal model $CBF(L)$, and the second ones for constructing the adversarial models $CBF(M)_1, \cdots, CBF(M)_m$. DARPA'99 provides additional traces with the specific purpose of serving to evaluate the sensors after trained, which were captured thorough two different weeks. Although the quality and usability of DARPA'99 at evaluating current proposals are questioned by the research community [19], it poses a functional standard that allows to compare novel contributions with previous proposals. With the purpose of evaluating the improvements of the proposal regarding APAP [17] and the previous contributions to the PAYL family [18], the dataset UCM 2011 was considered. It gathered real traffic traces monitored at the subnet of the Faculty of Computer Science at the University Complutense of Madrid (UCM). The traffic recording tasks were performed thorough the year 2011 at different time intervals and months. The resultant collection was labelled by the UCM Data Center. The normal traffic samples of UCM 2011 include normal network usage activities, among them: P2P exchanges, file transferences of several formats (.doc, .pdf, .mp3, .jpg, etc.) via SMTP, HTTP/HTTPS browsing, multimedia streaming, etc. As indicated in [18], the malicious binary contents were classified into 16 different groups of threats, including several malware families, DoD threats or privilege gain procedures. The experimentation stage took advantage of the mimicry attack generator described in [16]. This

Proposal	FPR (%)	TPR (%)
PAYL [31]	0.0	90.76
POSEIDON [4]	0.0	92.00
AnPDPP [27]	0.06	100.0
Anagram [30]	0.0	100.0
McPAD [21]	0.33	87.8
RePIDS [11]	0.67	99.33
APAP [17]	0.15	100
This proposal	0.01	100

Table 1: Results with non-obfuscated DARPA'99.

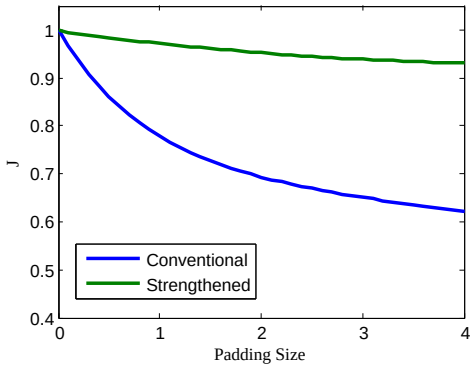

Figure 3: Padding size impact of mimicry (DARPA'99).

tool was originally developed for inserting padding actions on the sequences of malicious activities perpetrated by insider attackers. With the purpose of making the malicious traffic payloads gain similarity with the normal behaviors, the obfuscation tool had access to the $CBF(L)$ built at training stage. The selection of the padding sequences lied on the Probability Proportional-to-Size (PPS) stochastic sampling strategy [5]. In particular, the PPS algorithm begins with the extraction of the set of weights $P = \{p_1, \cdots, p_r\}$ corresponding with the r n-grams most frequent at $CBF(L)$, $0 < r$, where p_i is the number of occurrences of the binary sequence $b_i \in \{b_1, \cdots, b_r\}$.

4.2 Effectiveness with DARPA'99

As pointed out by [30] and ratified by [10, 17, 18], given the nature of the DARPA'99 dataset, a 3-gram sliding window proved to be the most suitable configuration for PAYL-inspired sensors. The proposal was configured accordingly. At this experiment, the adjustment parameter of the sensor was the confidence interval $\tau \in [-1, 1]$, thus being each payload p tagged as anomalous when the expression $GSC(p) < \tau$ is satisfied. As usual in previous comparisons, the obtained results are evaluated in the grounds of the best trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) estimated by the optimal Youden index J [2]. Table 1 displays the effectiveness of related publications and summarizes the obtained results. As indicated, the optimal calibration resulted in 100% hit rate and FPR=0.01% when J=0.9998. These results are close to those of the best sensor when operating under similar circumstances (ANAGRAM [30]), which proves the effectiveness of the proposal on the assumed experimental conditions. But despite these findings, the functional standardized use of DARPA'99 lacks of a way of proving the robustness of the sensors against adversarial threats, which in order to evaluate the rest of the enhancements of the proposal, leads to perform an additional experiment.

With the purpose of assessing the proposal strengthening against adversarial threats, this test considered a fixed confidence interval τ, in particular, that which displayed the optimal trade-off between TPR and FPR at the previous experiments. In this case, the calibration parameter measured was the variation of the incoherent labelling index ϕ, being hypothesized that the greater ϕ, the better true positive rate. The experiment assumed that every discovered labelling incoherency disguised a maliciously obfuscated payload. The variation of the best ϕ adjustment according to the Youden index is illustrated in Fig. 3, where it is observed that at both conventional and strengthened operational modes, the proposal is sensible to variations in the padding length. However, the strengthened approach significantly reduced the impact, being the observed accuracy decay saturated at approximately $AUC \approx 0.95$.

4.3 Effectiveness with UCM 2011

As summarized in Table 2, the best configuration (AUC=0.9900, J=0.9574) outperformed the best APAP (AUC=0.9136, J=0.8442) adjustment, but was very similar to APACS (AUC=0.9902, J=0.9339). However, since the proposal does not rely on alert correlation capabilities (as is the case of APACS), it entails a more efficient solution. On the other hand, APACS required a consistent and properly labeled knowledge-base, while our proposal provides fault tolerance at labelling errors of the training datasets.

The proposed was strengthened against adversarial attempts laying on mimicry, again proving its effectiveness when adopting the UCM 2011 dataset. This is illustrated in Figure 4, where variations of the Youden index based on the distribution of padding contents are displayed. At this experiment the confident interval τ was fixed to the optimal setting of the previous test, a 5-gram sliding windows was implemented, and the sensitiveness parameter was the adopted incoherent labelling index ϕ. In the same way that exposed in Figure 4, the strengthened version did not completely mitigate the mimicry impact, but as expected, it was significantly reduced. In this case the worst observed AUC approximated 0.9 which representatively enhanced the 0.7 achieved by a non-robust deployment. The saturation point approximated $AUC \approx 0.85$, while the original version was close to 0.6.

5 CONCLUSIONS

Thorough this paper, a novel approach for statistic malware detection on communication environments has been introduced. It was based on analyzing the traffic payloads looking for discordances regarding legitimate models previously build at training stage. The proposal adopted the basis of the PAYL sensors family, and extended the solutions APAP [17] and APACS [18], thus taking advantage of the n-gram methodology and the Bloom filter data structure paradigm. In contrast to most previous solutions, our research overcomes the traditional strengthening via randomization, by taking advantage of scoring the suspicious packet similarity between legitimate and previously built adversarial models. The proposal has

NIDS	FPR	TPR	AUC	J
This proposal	0.021	0.967	0.9900	0.9574
APAP [17]	0.080	0.947	0.9136	0.8442
APACS [18]	0.034	0.995	0.9902	0.9339

Table 2: Results with non-obfuscated UCM 2011.

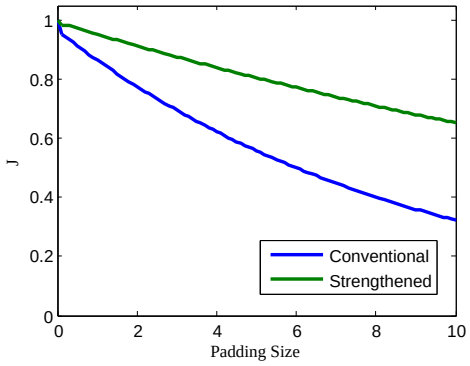

Figure 4: Padding size impact of mimicry (UCM 2011).

been evaluated according to the DARPA'99 evaluation methodology and the dataset UCM 2011, demonstrating high accuracy and similarity with the best contributions of the state-of-the-art. However, its effectiveness proved to outperform its antecessors when processing malicious payloads obfuscated by mimicry. But despite the extension of the detailed research, the discussion of some related interesting topics was postponed for future work, being in progress a research towards bringing robustness against more complex obfuscation paradigms (metamorphisms, fileless malware, etc.), as well as the proposal evaluation on alternative monitoring environments.

ACKNOWLEDGEMENTS

This work is funded by the European Commission Horizon 2020 Programme under grant agreement number 830892, as part of the project SPARTA (Special projects for advanced research and technology in Europe).

REFERENCES

[1] D. Ariu, R. Tronci, and G. Giacinto. 2011. HMMPayl: An intrusion detection system based on hidden Markov models. *Computers & Security* 30(4) (2011), 221–241.

[2] L.E. Bantis, C.T. Nakas, and B. Reiser. 2014. Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics* 70(1) (2014), 212–223.

[3] B. Biggio and F. Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 48 (2018), 317–331.

[4] D. Bolzoni, S. Etalle, P. Hartel, and E. Zambon. 2006. POSEIDON: a 2-tier Anomaly-Based Network Intrusion Detection System. In *Proceedings of the 4th IEEE International Workshop on Information Assurance (IWIA)*. London, United Kingdom, 144–156.

[5] E. Cohen and H. Kaplan. 2013. 16th International Workshop on Approximation, Randomization, and Combinatorial Optimization and 17th International Workshop, Randomization and Computation. In *Proceedings of the 15th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Berkeley, CA, USA, 452–467.

[6] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee. 2006. Polymorphic blending attacks. In *Proceedings of the 15th USENIX Security Symposium*. Vancouver, BC, Canada, 241–256.

[7] P. Garcia-Teodoro, J.E. Diaz-Verdejo, J.E. Tapiador, and R. Salazar-Hernandez. 2015. Automatic generation of HTTP intrusion signatures by selective identification of anomalies. *Computers & Security* 55 (2015), 159–174.

[8] R. Green, I. Staffell, and N. Vasilakos. 2014. Divide and Conquer? k-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System. *IEEE Transactions on Engineering Management* 61(2) (2014), 251–260.

[9] D. Hadziosmanovik, L. Simionato, D. Bolzoni, E. Zambon, and S. Etalle. 2012. N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols. In *Proceedings of the 15th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Amsterdam, The Netherlands, 59–81.

[10] D. Hadziosmanovik, L. Simionato, D. Bolzoni, E. Zambon, and S. Etalle. 2012. N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols. In *Proceedings of the 15th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Amsterdam, The Netherlands, 59–81.

[11] A. Jamdagni, Z. Tan, X. He, P. Nanda, and R.P. Liu. 2013. RePIDS: A multi tier Realtime Payload-based Intrusion Detection System. *Computer Networks* 57 (2013), 511–824.

[12] X. Jin, B. Cui, D. Li, Z. Cheng, and C. Yin. 2018. An improved payload-based anomaly detector for web applications. *Journal of Network and Computer Applications* 116 (2018), 111–116.

[13] T. Jonathon, J. Somesh, and B.P. Miller. 2006. Automated Discovery of Mimicry Attacks. In *Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Hamburg, Germany, 41–60.

[14] A. Karami. 2018. An anomaly-based intrusion detection system in presence of bening outliers with visualization capabilities. *Expert Systems with Applications* 108 (2018), 36–60.

[15] R. Lippmann, J.W. Haines, D.J. Fried, J. Korba, and K. Das. 2000. The 1999 DARPA off-line intrusion detection evaluation. *Computer Networks* 34(4) (2000), 579–595.

[16] J. Maestre Vidal, A.L.S. Orozco, and L.J.G. Villalba. 2016. Online masquerade detection resistant to mimicry. *Expert Systems with Applications* 61 (2016), 162–180.

[17] J. Maestre Vidal, A.L.S. Orozco, and L.J.G. Villalba. 2017. Advanced Payload Analyzer Preprocessor. *Future Generation Computer Systems* 76 (2017), 474–485.

[18] J. Maestre Vidal, A.L.S. Orozco, and L.J.G. Villalba. 2017. Alert correlation framework for malware detection by anomaly-based packet payload analysis. *Journal of Network and Computer Applications* 97 (2017), 11–22.

[19] J. McHugh. 2000. Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security* 3(4) (2000), 262–294.

[20] S. Pastrana, A. Orfila, J.E. Tapiador, and P. Peris-Lopez. 2014. Randomized Anagram revisited. *Journal of Network and Computer Applications* 21 (2014), 182–186.

[21] R. Perdisci, D. Ariu, P. Fogla, G. Giacinto, and W. Lee. 2009. McPAD: A Multiple Classifier System for Accurate Payload-Based Anomaly Detection. *Computer Networks* 53(6) (2009), 864–881.

[22] O. Rottenstreich and I. Keslassy. 2015. The Bloom paradox: when not to use a Bloom filter. *IEEE/ACM Transactions on Networking* 23(3) (2015), 703–716.

[23] J. Shana and T. Venkatachalam. 2014. An Improved Method for Counting Frequent Itemsets Using Bloom Filter. *Procedia Computer Science* 47 (2014), 84–91.

[24] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and Chanona-Hernández L. 2014. Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications* 41(3) (2014), 853–860.

[25] M. Swarnkar and N. Hubballi. 2016. OCPAD: One class Naive Bayes classifier for payload based anomaly detection. *Expert Systems with Applications* 64 (2016), 330–339.

[26] J.E. Tapiador and J.A. Clark. 2010. Information-Theoretic Detection of Masquerade Mimicry Attacks. In *Proceedings of the 4th International Conference on Network and System Security*. Melbourne, VIC, Australia.

[27] S.A. Thorat, A.K. Khandelwal, B. Bruhadeshwar, and K. Kishore. 2009. Anomalous packet detection using partitioned payload. *Journal of Information Assurance and Security* 3(3) (2009), 195–220.

[28] A. Viswanathan, K. Tan, and C. Neuman. 2013. Deconstructing the Assessment of Anomaly-based Intrusion Detectors. In *Proceedings of the 16th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Rodney Bay, St. Lucia, 286–306.

[29] K. Wang, G. Cretu, and S.J. Stolfo. 2005. Anomalous Payload-based Worm Detection and Signature Generation. In *Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Seattle, WA, USA, 227–246.

[30] K. Wang, J.J. Parekh, and S.J. Stolfo. 2006. Anagram: A Content Anomaly Detector Resistant to Mimicry Attack. In *Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Hamburg, Germany, 226–248.

[31] K. Wang and S.J. Stolfo. 2004. Anomalous Payload-based Network Intrusion Detection. In *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID)*. Sophia Antipolis, France, 203–222.