International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

A Literary Review of Pattern Matching Techniques in Network Intrusion Detection

Nitin Venkatesh[1], Prof. Pradnya Kashikar[2]
[1]Student, Birla Institute of Technology and Science, Pilani - India
[2]Adjunct Faculty, Birla Institute of Technology and Science, Pilani - India

ABSTRACT: With the exponential growth in devices and services being added to networks, we are also witnessing an increase in the volume and complexity of threats, urging an increased efficiency in network intrusion detection systems which primarily rely on pattern matching to identify malicious activity on the network. In this literary review of pattern matching techniques in network intrusion detection, we explore the limitations and the research carried out in both signature-based and anomaly-based intrusion detection systems to overcome them. It focuses on the performance improvements in signature-based intrusion detection systems achieved through methodologies and technologies like regular expressions, Hyperscan, RE2, Flashtext, a generalized Aho-Corasick algorithm, usage of Bloom filters and payload sampling. It also covers the usage of machine learning techniques, including genetic algorithms, Support Vector Machines (SVM) and Improved Self-Adaptive Bayesian Algorithm (ISABA), which are used to detect anomalous behavior and identify potential threats in a network in anomaly-based network intrusion detection to assist the security analysts carry out their job functions. Additionally, this review explores the integration of the MITRE ATT&CK framework and Security Information and Event Management (SIEM) systems in network intrusion detection as this framework provides a structured and standardized approach for analyzing the tactics and techniques used by attackers to classify them, while SIEM systems enable the correlation of threat activity across multiple sources, allowing for a more comprehensive and accurate view of the network security. Overall, this literary review provides insights into the state-of-the-art techniques and frameworks used in Network Intrusion Detection based on Pattern Matching, highlighting the significant improvements in performance and detection capabilities.

KEYWORDS: Intrusion Detection, Machine Learning, Network Security, Pattern Matching, Regular Expressions.

INTRODUCTION

With the ongoing digital revolution, we are witnessing a huge addition of devices to networks, both public and private, and the rate of data exchange including sensitive information as more services and businesses are adopting the digital trend has grown manifold requiring organizations to be cognizant of all the activity on their network and ensure that their network is secure. Network security has become paramount for organizations as the number of threats in cyberspace has grown exponentially along with the digital revolution.

This is evident from some high-profile network intrusion incidents in the recent past like the SolarWinds supply chain attack (2020) where the attackers compromised the build system and injected malicious code into a software update to all customers of the Orion product of SolarWinds resulting in a number of government as well as private organizations being impacted and the massive data breaches of Equifax (2017) and Target (2013) which resulted in the personal information of millions of customers being leaked leaving them vulnerable to identity theft and targeted spear-phishing attacks.

In this literary review, we aim to summarize some of the popular techniques that could be used in pattern matching for network intrusion by signature-based systems and anomaly-based systems. The review aims to highlight the strengths and weaknesses of the techniques discussed, thus trying to provide a fair overview of the past research conducted in the field of pattern matching with a network intrusion detection context.

OVERVIEW OF NETWORK INTRUSION DETECTION SYSTEMS

Network intrusion detection systems are available as a hardware or software appliance and could be present out-of-band or inline based on the requirements of the organizations and its security team. They can be classified majorly based on the techniques they use as

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

Signature-based: based on rules specifying the pattern explicitly.
Anomaly-based: based on a deviation of behavior from the standard behavior the baselines for which could be explicitly defined or modeled by the system itself based on historical network activity.
Hybrid: a combination of both signature and anomaly based intrusion detection system employing the best of both worlds.
Signature-based systems form the cornerstone of intrusion detection systems just like in most other security systems and the reason for it being the sheer adaptability and accuracy that it wields. It becomes easy to add Indicators of Compromise for new threats as new signatures to the existing rule set providing immediate protection against the known vectors. It is common to see auto-updation of rules provided by trusted third-party vendors to secure against new and emerging threats.
Signature-based systems do have some limitations when it comes to performance as it depends on various factors including but not limited to the rules, the rule checking engine / algorithm and the data being inspected itself and hence can vary drastically in different environments or also in the same environments with different data / traffic. A major limitation of signature-based systems are that it can only defend against known threats that it has a signature for.

PERFORMANCE IMPROVEMENT OF PATTERN MATCHING IN SIGNATURE-BASED INTRUSION DETECTION SYSTEMS

Regular expressions (or regex) are the de-facto standard to match strings and their variations as patterns in most security-related software mostly due to the flexibility it offers, but that flexibility comes at a price of performance. Generally speaking, regexes are slower when compared to fixed string matches and the performance could be impacted even more significantly depending on the way the regex search pattern is written and also the regex engine used. When inspecting large amounts of traffic, the cost of using regexes becomes quite high to the point where it sometimes could end up being not useful.
Most signatures used in signature-based systems have certain fixed strings in the patterns they are searching for. Since fixed-string matches are faster than compiling and searching with a regex for the same, new methods and algorithms have been researched to provide faster alternatives. The Aho-Corasick algorithm [1] has influenced most pattern matching algorithms since its introduction in 1975 and till this day remains one of the fastest ways to perform a text search. FlashText [2], an algorithm inspired by the Aho-Corasick algorithm, was designed as a fast search and replacement algorithm to replace long regexes consisting of fixed strings with word-boundary on either sides, for example: \b(ssh|ftp|telnet|http|https|smtp)\b so this algorithm would be useful in scenarios where we only need to match the search terms as a whole and not when it appears as a substring, for example, it would not match a term like openssh, sshd or httpd. The performance gain is remarkable when using FlashText as a replacement for regexes when the number of search terms are greater than 100 and surrounded by word-boundaries. FlashText performs with a time-complexity that is dependent on the number of characters in the document being searched rather than the number of terms being searched. So, for a document of size N (characters) and a dictionary of M keywords, the time complexity will be O(N) which bodes well against regex which requires a time complexity of O(M*N) for the same.
Although FlashText is great to search for complete words in documents, using at least some simple regular expression constructions provides a lot more flexibility which would greatly improve the coverage offered by the signatures. A generalized Aho-Corasick algorithm was researched [3] to construct a finite state pattern matching machine that allows for usage of operators like ? (symbol maybe be present or not), * (symbol may occur any number of times consecutively), {min, max} (symbol lies in the range between min and max consecutive occurrences). Although the research paper uses signatures found in ClamAV (an anti-virus product) to establish the efficiency of the algorithm, the same can be extrapolated to Intrusion Detection systems where simple regexes are used.
Another similar research [4] explored splitting the Finite State Machine into smaller ones in order to optimize Aho-Corasick algorithm and exploit the domain specific characteristics of intrusion detection where the headers of the incoming packets are checked and categorized into groups before activating the corresponding Finite State Machine to scan the payload of the packet.
There has also been research [5] around using Bloom filters combined with multi-pattern search (using Aho-Corasick algorithm) where a longest prefix match is done preliminarily, during which a majority of the memory accesses are suppressed as Bloom filters are used, after which the longer string matching, if required, is done by coupling it with the Aho-Corasick algorithm. An idea of scanning only parts of the payload instead of the entire traffic data in deep-packet inspection has also been researched [6] where the amount of text to be searched for is greatly reduced which improves the performance significantly. In this research, the method proposed involves using two Deterministic Finite Automata (DFAs) - a “sampled” DFA, used to perform fast search and exclude

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

most of the non-malicious traffic and a "reverse" DFA to perform a more accurate processing in order to confirm a match and reduce False Positives that occur due to pattern matching on sampled data. This research found that using the sampling technique described improves the performance by 350% over standard DFAs. The above discussed techniques allow for scalability and speed, making them good candidates for use in network intrusion detection systems.

The complexity and features of the PCRE (Perl-Compatible Regular Expression) engine are unparalleled and allow for complex search patterns to be written which is absolutely required by the security professionals to defend the network and devices on them against potential threats. The PCRE engine although full-featured is resource hungry and doesn't perform as well with larger datasets, thus inspiring researchers to explore more efficient regex engines with most if not all the features offered by PCRE but be scalable and effortlessly work on large datasets and large number of patterns. Google's RE2 and Intel's Hyperscan engines are examples of the outcomes of such research [7][8]. RE2 offers the option of a slightly modified or limited implementation of the PCRE syntax whereas Hyperscan follows the PCRE syntax exactly. Hyperscan also utilizes the hardware and software optimizations to maximize the performance of the pattern matching engine and guarantees non-exponential scanning time and immunity from Re-DOS, a common problem observed with regexes when searching through crafted data. Apart from its multi-pattern searching capability that assures performance as the number of terms grow in size, it also features a "streaming" mode that allows for "cross-packet" inspection as stream writes are written sequentially to a logical stream and regex matches are detected for cases that cross stream write boundaries, making it ideal for use in network intrusion detection systems. Although Hyperscan is performant than RE2 in most cases, RE2 has its niche and can be deployed in most environments whereas Hyperscan only has limited hardware compatibility as it requires processors to be able to use the Intel Streaming SIMD Extensions 3 instruction set as a minimum requirement, which might not be available in all environments.

Table I. Performance Comparison with PCRE, PCRE2, RE2 and Hyperscan (in Single and Multi Match Modes) for Snort Talos (1,300 regexes) and Suricata (2,800 regexes) rulesets with Real Web Traffic Trace. Numbers are in Seconds.[Source: [8], Table 5]

Ruleset	PCRE	PCRE2	RE2-s	Hyperscan-s	RE2-m	Hyperscan-m
Talos	6,942	394	1,777	173	29	2.15
ET-Open	12,800	913	4,696	516	1,116	133

PATTERN MATCHING IN ANOMALY-BASED INTRUSION DETECTION SYSTEMS

Signature-based systems can assist in defending against known attacks and attack patterns, but zero-day attacks are impossible to detect unless signatures are created for them and hence they remain invisible on the network until then, a technique commonly used by Advanced Persistent Threats (APTs). To counter such threats, the normal traffic patterns need to be known in order to distinguish it from malicious traffic. Anomaly-based intrusion detection systems require a baselining of the normal traffic which can be provided by an expert with the specific domain knowledge or letting the system figure out what the baseline is by itself by studying normal traffic patterns and then defining the criteria or rules for anomalies. Since expert human based baselining and analysis purely depends on the skill and domain knowledge of the expert designing the system, we cover the machine learning approaches employed to identify and classify anomalous network traffic in this literary review.

Research has been conducted on generating rules using genetic algorithms and decision trees to classify network traffic [9]. The system developed by this research layers onto existing network-based Intrusion Detection Systems to monitor traffic and correlate numerous intrusion patterns to produce rules for compilation into the existing IDS.

Later research explored the use of Support Vector Machines (SVMs) for classification of network traffic since it has a better learning ability for small samples [10]. The features of SVMs like insensitiveness to dimensionality of input data and the self-learning ability offered by continuous correction of various parameters with an increase in the training data have made it a popular area of research as well as use by Intrusion Detection Systems employing machine learning learning models. Another research explored the use of a self-adaptive Bayesian algorithm to classify network traffic [11]. The Bayesian model provides a probabilistic approach to classification which makes it a good model to predict the class of an unknown sample. The Improved Self-Adaptive Bayesian Algorithm (ISABA) described in the paper works by initializing the weights of each example to 1.0 and estimates the prior

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

probability of each class, updating the probability distribution based on new data it encounters using a Bayesian approach while taking into account both the new data as well as the prior distribution. The ISABA algorithm was compared with other machine learning models like SVMs, Neural Networks, Genetic Algorithms and Naive-Bayes Classifier using 494020 data samples for training and 311028 samples for testing, where ISABA outperformed the detection rate by other models for different types of malicious network traffic and outperformed them.

Table II. Performance of Naive-Bayes vs ISABA - detection rate vs false-positive rate. [Source: [11], Table 3]

	Normal	**Probe**	**DoS**	**U2R**	**R2L**
Naive Bayesian Classifier (DR %)	99.25	99.13	99.69	64	99.11
Naive Bayesian Classifier (FP %)	0.08	0.45	0.04	0.14	8.02
Improved Self Adaptive Bayesian Algorithm (DR %)	99.62	99.22	99.49	99.17	99.15
Improved Self Adaptive Bayesian Algorithm (FP %)	0.05	0.36	0.03	0.10	6.91

Table III. Performance of various machine learning models. [Source: [11], Table 5]

	SVM	**NN**	**GA**	**NB**	**ISABA**
Normal	99.4	99.6	99.3	99.55	99.82
Probe	89.2	92.7	98.46	99.43	99.72
DoS	94.7	97.5	99.57	99.69	99.49
U2R	71.4	48	99.22	64	99.47
R2L	87.2	98	98.54	99.11	99.35

A recent review [12] conducted on Intrusion Detection research trends found that the most implemented models were k-Nearest Neighbours - 7%, Random Forests - 7%, Naive Bayes - 15% , Decision Trees - 17%, Neural Networks - 20% and Support Vector Machines - 34% . The review suggested future research on ensemble techniques combining multiple models and boosting algorithms and combined feature selection algorithms to improve accuracy of machine learning techniques in intrusion detection.

As we have a huge number of devices on the network, many of them specialized security software or appliances, like Host-based Intrusion Detection Systems (HIDS), endpoint-security solutions, OS event logs, Data Loss Prevention (DLP), Web Application Firewalls (WAF), etc., leveraging the data and logs from them is crucial to better discern patterns that might not always be visible to the traditional IDS alone. Security Information and Event Management (SIEM) systems help digest the logs from multiple sources and allow for correlation of data to raise alerts and incidents for analysts to investigate and take remediative actions if necessary. Since there is voluminous data from multiple sources, research [13] has been made on applying deep learning techniques for enhanced threat detection based on event profiling that generates alerts for the analysts by comparing long-term security data. Research has also been conducted on integrating malware detection techniques like function hooking sending data to the SIEM [14] that would help in uncovering the Advanced Persistent Threat (APT) activity which is very hard to detect since the intruder is highly

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

skilled and leverages zero-day exploits that are impossible to detect without behavior monitoring. The ATT&CK framework [15] provided by the MITRE corporation enables us to understand the tactics, techniques and procedures used by various threats thus enhancing the event profile to provide more insights into threat activity and help identify them as belonging to a particular threat instead of being disparate events.

CONCLUSION

This literary review focused on prevalent network intrusion detection techniques in both signature-based systems as well as anomaly-based systems. The main advantage of signature-based systems is the speed and accuracy, which behavior-based systems cannot offer at this point in time. But the disadvantage that a signature-based system has is that it can only detect known malicious activity that signatures are created for and hence novel, zero-day exploits and threat activity like that of Advanced Persistent Threats (APTs) leveraging them are invisible to it.

Ideas of replacing regexes with fixed-strings where applicable, for faster searching with algorithms like FlashText, application of sampling techniques and the use of regex engines like Hyperscan have proven to be effective and scalable, boosting the performance of the traditional signature-based systems.

Using techniques leveraging machine learning models like Support Vector Machines, Genetic Algorithms, Improved Self-Adaptive Bayesian Algorithms have shown a high detection rate with a low false-positive rate making them good choices for anomaly-based systems. However, more research into ensembling techniques combining multiple models shows promise of improved performance. Rather than a single appliance trying to detect intrusions based on network activity, leveraging the numerous devices (security devices, workstations, servers, etc.) and their logs to correlate activity on the network provides better visibility to the analyst. The use of the Security Incidents and Events Monitoring (SIEM) systems can correlate the information to provide alerts and incidents to the analysts to investigate and take any remediative action if required. The application of artificial intelligence to provide timely alerts is gaining momentum and we can expect most of the triage work currently carried out by analysts to be handled by an AI allowing the analyst to focus on critical alerts and incidents. The log data could also be used in conjunction with the ATT&CK framework to conduct threat hunting, understand the techniques, tactics and procedures of an adversary and defend against them.

Based on the literary review carried out, we could plan to next work on research and experiment on the effectiveness of procedures using artificial intelligence techniques in threat-hunting by leveraging correlated data from multiple data sources layered on frameworks like the MITRE ATT&CK framework to discover zero-day exploits, Advanced Persistent Threats and unknown malicious activity, and converting them into signatures for use in signature-based intrusion detection systems, thus leveraging the best of both worlds in a hybrid model to produce an effective intrusion detection system.

In this literary review, we have summarized the research and techniques used popularly in pattern matching that could be used in the context of network intrusion detection. The review explored the use of pattern matching techniques in signature-based systems using fixed strings, regular expressions and the advancements made to improve their performance suitable to the current trends and needs of analyzing voluminous data. The review also briefly explores the research in machine learning techniques to analyze malicious network traffic and its increasing relevance to defend against network intrusions as large amounts of data from various network endpoints and security applications need to be correlated to understand the network activity better to improve and maintain the security posture of organizations against unknown and emerging threats.

REFERENCES

1. Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: an aid to bibliographic search. Commun. ACM 18, 6 (June 1975), 333–340. https://doi.org/10.1145/360825.360855.
2. V. Singh, ‘Replace or Retrieve Keywords In Documents at Scale’, arXiv [cs.DS]. 2017.
3. T. -H. Lee, "Generalized Aho-Corasick Algorithm for Signature Based Anti-Virus Applications," 2007 16th International
4. Conference on Computer Communications and Networks, Honolulu, HI, USA, 2007, pp. 792-797, doi: 10.1109/ICCCN.2007.4317914.
5. V. Dimopoulos, I. Papaefstathiou and D. Pnevmatikatos, "A Memory-Efficient Reconfigurable Aho-Corasick FSM Implementation for Intrusion Detection Systems," 2007 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, Samos, Greece, 2007, pp. 186-193, doi: 10.1109/ICSAMOS.2007.4285750.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 09 September 2023
DOI: 10.47191/ijcsrr/V6-i9-40, Impact Factor: 6.789
IJCSRR @ 2023

6. S. Dharmapurikar and J. W. Lockwood, "Fast and Scalable Pattern Matching for Network Intrusion Detection Systems," in IEEE Journal on Selected Areas in Communications, vol. 24, no. 10, pp.1781-1792, Oct. 2006, doi: 10.1109/JSAC.2006.877131.
7. D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procissi and F. Vitucci, "Sampling Techniques to Accelerate Pattern Matching in Network Intrusion Detection Systems," 2010 IEEE International Conference on Communications, Cape Town, South Africa, 2010, pp. 1-5, doi: 10.1109/ICC.2010.5501751.
8. Russ Cox, "Regular Expression Matching in the Wild", 2010 (available at https://swtch.com/~rsc/regexp/regexp3.html).
9. Xiang Wang, Yang Hong, Harry Chang, KyoungSoo Park, Geoff Langdale, Jiayu Hu, and Heqing Zhu. 2019. Hyperscan: a fast multi-pattern regex matcher for modern CPUs. In Proceedings of the 16th USENIX Conference on Networked Systems Design and Implementation (NSDI'19). USENIX Association, USA, 631–648.
10. C. Sinclair, L. Pierce and S. Matzner, "An application of machine learning to network intrusion detection," Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99), Phoenix, AZ, USA, 1999, pp. 371-377, doi: 10.1109/CSAC.1999.816048.
11. X. Bao, T. Xu and H. Hou, "Network Intrusion Detection Based on Support Vector Machine," 2009 International Conference on Management and Service Science, Beijing, China, 2009, pp. 1-4, doi: 10.1109/ICMSS.2009.5304051.
12. Farid, Dewan & Zahidur Rahman, Mohammad. (2010). Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm. Journal of Computers. 5. 10.4304/jcp.5.1.23-31.
13. Amarudin, R. Ferdiana and Widyawan, "A Systematic Literature Review of Intrusion Detection System for Network Security: Research Trends, Datasets and Methods," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2020, pp. 1-6, doi: 10.1109/ICICoS51170.2020.9299068.
14. J. Lee, J. Kim, I. Kim and K. Han, "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles," in IEEE Access, vol. 7, pp. 165607-165626, 2019, doi: 10.1109/ACCESS.2019.2953095.
15. N. A. S. Mirza, H. Abbas, F. A. Khan and J. Al Muhtadi, "Anticipating Advanced Persistent Threat (APT) countermeasures using collaborative security mechanisms," 2014 International Symposium on Biometrics and Security Technologies (ISBAST), Kuala Lumpur, Malaysia, 2014, pp. 129-132, doi: 10.1109/ISBAST.2014.7013108.
16. MITRE ATT&CK. (2020). Design and Philosophy. (available at https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf).

Cite this Article: Nitin Venkatesh, Prof. Pradnya Kashikar (2023). A Literary Review of Pattern Matching Techniques in Network Intrusion Detection. International Journal of Current Science Research and Review, 6(9), 6491-6496