International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 2
ISSN: 2088-5334

Probabilistic Analysis of Random Check Intrusion Detection System

Firuz Kamalov [a], Sherif Moussa [a,*], Gandeva Bayu Satrya [a]

[a] *Faculty of Engineering, Canadian University Dubai, Dubai, United Arab Emirates*
*Corresponding author: *firuz@cud.ac.ae*

***Abstract*—The ubiquitous adoption of network-based technologies has left organizations vulnerable to malicious attacks. It has become vital to have effective intrusion detection systems (IDS) that protect the network from attacks. In this paper, we study the intrusion detection problem through the lens of probability theory. We consider a situation where a network receives random malicious signals at discrete time instances, and an IDS attempts to capture these signals via a random check process. We aim to develop a probabilistic framework for intrusion detection under the given scenario. Concretely, we calculate the detection rate of a network attack by an IDS and determine the expected number of detections. We perform extensive theoretical and experimental analyses of the problem. The results presented in this paper would be helpful tools for designing and analyzing intrusion detection systems. We propose a probabilistic framework that could be useful for IDS experts; for a network-based IDS that monitors in real-time, analyzing the entire traffic flow can be computationally expensive. By probabilistically sampling only a fraction of the network traffic, the IDS can still perform its task effectively while reducing the computational cost. However, checking only a fraction of the traffic increases the possibility of missing an attack. This research can help IDS designers achieve appropriate detection rates while maintaining a low false alarm rate. The groundwork laid out in this paper could be used for future research on understanding the probabilities related to intrusion detection.**

***Keywords*— Intrusion detection system; detection rate; uniform distribution; Poisson distribution.**

Manuscript received 7 Mar. 2023; revised 9 Jan. 2024; accepted 7 Mar. 2024. Date of publication 30 Apr. 2024.

I. INTRODUCTION

The rapid progress in network-based technologies has increased the threat of spammers, attackers, and criminal enterprises. The economic loss from malicious network attacks is estimated to be hundreds of billions of dollars. Therefore, intrusion detection systems (IDS) have become as crucial as ever. In this paper, we analyze the probabilistic aspects of intrusion detection. Concretely, we consider a scenario where only a fraction of the network signals is checked by an IDS due to the associated costs. We analyze the likelihood of detecting a malicious network signal by an IDS performing uniformly random checks. The theoretical framework presented in this paper will help design and analyze intrusion detection systems. Comprehending the probabilities of intrusion detection based on the number of attacks and security checks would allow us to calibrate IDS to achieve a high detection rate while maintaining a low false alarm rate.

Scanning the network for potential malware requires considerable time and computational resources. The time needed to process a signal by the IDS slows the network traffic. High-volume traffic also requires extensive computational capabilities. Given the above considerations, many IT departments choose to forego a full scan of the network. As a result, a malicious signal can go undetected. There is a relationship between the amount of scanned traffic and the probability of intrusion detection. As the amount of scanned traffic signals increases, so does the likelihood of detecting a malicious signal. Understanding the exact nature of this relationship requires a probabilistic point of view.

Our goal in this paper is to analyze the probabilities related to intrusion detection. We calculate the probability of an IDS detecting a malicious signal under the assumption of a uniform distribution of attacks and security checks. We also calculate the expected number of detected malicious signals. Numerical experiments support our theoretical findings.

Technological advances such as cloud services and IoT have dramatically increased the demand for secured network connections. Cyberattacks pose a significant risk, threatening all connected activities, such as governmental, educational, business, and healthcare systems [1]. The abundant knowledge and resources today have reduced barriers to entry for hackers. Attackers continuously exploit network vulnerabilities to steal critical information or deny the provided services. For example, a tiny IoT device can be compromised to run an attack and disrupt the operations of large cloud networks, manufacturing facilities, power plants, and others [2], [3]. The learning curve to develop a sophisticated attack has been shortened due to the considerable knowledge base available on the internet. The negative consequences of network attacks have risen dramatically. The financial loss due to global cybercrime is estimated to be as high as 6 trillion dollars by 2021 [4].

Concerns over protecting vital computer networks have accelerated network and online security research. Cybersecurity has become a significant research area. Its primary aim is to develop techniques that protect computer networks and overcome cybercrimes. While conventional security techniques such as firewalls, encryption, user authentication, and access control lists play an essential role in protecting data on the network, they are not enough to provide complete protection against cyber threats [5]. The increased complexity of attacks requires constant real-time monitoring and detection of new security breaches.

An intrusion detection system (IDS) is often employed to address this issue. An IDS is designed to monitor and analyze the inbound connections and activities of the computer network and to look for any unauthorized connections or illegal activities that violate the standard security practice or network security policies [6]. IDS gathers, logs, detects, and prevents any security breaches. High detection rates and low negative alarms are essential factors for IDS systems. IDS could be implemented as hardware or software to automatize the detection of intruders, monitor traffic for illegal activities, and send alarms to network administrators. An intrusion prevention system (IPS) is another popular network protection tool. It is hardware or software that prevents illegal access to network resources. Today, security devices combine both intrusion detection and prevention functionalities. Intrusion detection and prevention systems (IDPS) detect, record, and prevent attack incidents and send alarms to network administrators [6].

Many network-based IDS function actively, processing incoming traffic in real-time. Given a high volume of network traffic, an IDS might struggle to handle all the incoming packets. To address this issue, only a fraction of the packets may be chosen for processing. However, checking only a fraction of the traffic introduces a probability of missing an attack.

Our paper attempts to calculate the probability of catching an attack given the number of malicious signals sent by an attacker and the number of checks performed by the IDS. We believe that understanding the relationship between the probability of catching an attack, the number of attacks, and the number of checks by the IDS will help design/fine-tune the IDS to achieve desired outcomes. Despite simplified assumptions on the attack behavior and IDS checking procedure, we believe the analysis presented in the paper helps advance our understanding of the problem.

Our paper is organized as follows: In Section 2, we review the existing literature on intrusion detection and related probabilistic models. In addition to Section 2, we state and prove our main results. In Section 3, we present the results of experiments that were carried out to validate our theoretical predictions. Section 4 concludes the paper.

II. Material and Method

A. Literature Review

This section discusses the existing literature on intrusion detection and the related probability models. Intrusion detection systems can detect network intrusions. They generate an alarm or log the results. IDS can be classified into three categories based on the detection method used in the system: signature-based, anomaly-based, and hybrid detection [7]. Signature-based, also known as misuse, detection uses the knowledge from previous attacks to create a unique signature for each attack and store it in a database. Knowledge-based techniques can achieve high levels of accuracy by leveraging the knowledge accumulated from past attacks. Since the intrusion alarm is generated only if the attack signature matches the one stored in the database, the chance of having a false positive is meager. However, if the attack signature is not found in the signature database, the method will fail to detect the possible intrusion. Such a scenario is known as a zero-day attack [8].

The second approach is the anomaly detection method. In this technique, normal network activities are modeled and considered as an operation baseline. Any deviation from this baseline is regarded as an attack. In other words, abnormal behavior could be identified by comparing incoming traffic during the monitoring time to a predefined traffic profile. Consequently, the chance of detecting zero-day attacks is very high using anomaly-detection IDS. The baseline model is constructed using the previously collected data from network hosts, users, and connections during regular operation [9]. The baseline model can be updated offline or online. The model remains static and unchanged for the offline update until the IDS generates a new request to create an updated profile. On the other hand, online update techniques include the statistical analysis of dynamic thresholding for some network attributes [10].

The growth in computing power has made machine learning and data mining techniques top-rated detection tools. Machine learning techniques used in intrusion detection include deep learning [11], genetic algorithms [12], and neural networks [13]. Data mining techniques include feature selection [14,15,16] and outlier detection [17]. Since there are many situations where the network deviates from regular operation, such as faulty devices or protocols, anomaly-based IDS might consider activities related to those situations as an attack. As a result, the system may suffer from a high false positive rate. Additionally, creating a profile for regular traffic requires a large amount of time and data resources to train the anomaly-based IDS.

The third approach is the hybrid detection method, which is a combination of the above two methods used to alleviate the weaknesses of the signature-based and anomaly techniques. To reduce zero-day attacks and big false positive problems, multiple algorithms must be processed concurrently to decide the anomaly of an event. At the same time, the algorithm must match the event signature to the previously recorded attack [18]. Depending on its location within the network, an IDS can be classified as a host- or network-based system. A host-based IDS is configured to monitor single-host activities to identify malicious attacks that network-based systems cannot detect. It uses a database of historical behavior to detect malicious activities. Host-based IDS are designed to discover attacks generated within the network but suffer from slow response to real-time attacks [19].

On the other hand, a network-based IDS uses multiple sensors to monitor activities related to the entire network. It can identify attacks affecting multiple hosts [20]. Unlike the host-based IDS, the network-based IDS are considered active systems due to their real-time examination of all packets in the incoming traffic flow. Hence, the packet processing rate is a significant challenge for network-based IDS. The detection system must process data as fast as the incoming packet rate.

Achieving an adequate performance speed might be very difficult for an extensive network. One approach to address this issue is considering a scenario where an IDS checks only a fraction of the traffic flow. In other words, the system will monitor only a probabilistically sampled fraction of the total flow [21]. A comprehensive literature exists on sampling methods applied to traffic classification [22]. A thorough comparison is reported by [23]. As noted in [24], sampling techniques can be classified into four categories: packet, flow, smart, and selective sampling. It is known that while packet sampling outperforms flow sampling, smart and selective techniques perform better than two. In [25], the authors propose an adaptive, feature-aware statistical sampling technique and compare it formally and empirically with other known sampling techniques - random flow and selective sampling. In [26], the authors show that intelligent flow sampling can improve anomaly detection performance. Their entropy-based method combines opportunistic and preferential data sampling to magnify anomalies in the sampled data, improving their detection. In this paper, we focus on packet sampling.

The authors in [27] analyzed the effect of random packet sampling on anomaly detection. They introduced anomaly detection metrics and applied their methodology to blaster worm data. They proved that sampling does not affect the anomaly size measured in several bytes or packets. Here, we look at the same problem from a mathematical perspective. We propose a mathematically rigorous approach to address the issue of intrusion detection. We prove that a simple random sampling mechanism can guarantee a robust detection rate. Our paper provides the mathematical framework and expressions for anomaly detection probability. Due to the limited computational resources, the statistical model can help IDS designers establish trade-offs between checking all packet flow and a subset of the flow. Molding the detection sampling will enable network administrators to make informative decisions considering the trade-offs between resource limitations and IDS accuracy.

B. Intrusion Detection Rate

In this section, we supply our main results and the corresponding proofs. We will calculate the theoretical probability of detecting a malicious signal by an IDS that checks the incoming signal at random points in time. We make certain simplifications to network attack and detection scenarios to allow for mathematical investigation. Consider a setting where a system receives malicious signals over a fixed interval. Suppose an IDS is designed to check periodically whether an incoming signal is malicious. Our goal is to compute the probability of detecting an incoming malicious signal. To formalize the problem, we make the following assumptions,

- Let I denote a fixed time interval. Interval I is divided into n discrete instances at which signals can be sent. We denote the time instances by $\{t_i\}_{i=1}^{n}$. The signals and checks can occur at any of these instances. For example, if the time interval I has length is 1 minute and $n = 60$, then the instances will be one second apart.
- Let k be the number of malicious signals that are sent during the time interval I. We assume that the malicious signals are sent randomly and independently.
- Let m be the number of times the IDS performs its checks. We assume that IDS checks occur randomly and independently.

We can model the above situation as an experiment based on random selection of a sequence of numbered balls as shown in Fig. 1. Suppose we have a set of balls numbered 1 through n. The balls represent the instances in the time interval I. Suppose that k of the n balls is also marked. The marked balls represent the time instances at which malicious attacks occur. We then randomly choose m balls without replacement from the collection of n balls. The selected balls represent the instances at which the IDS performs a check. Since the selections correspond to the checks performed by the IDS, they must be done in an ordered fashion. For example, having a choice (3, 5, 7) would be possible, but not (7, 3, 5). The second sequence would mean that the first check was done at instance number 7, and the second check was done at instance number 3, which is physically not possible. In other words, each selected combination of balls must be numbered in increasing order. The temporal considerations introduce another dimension to the problem. We must take great care in our analysis to satisfy the temporal order of the choices.

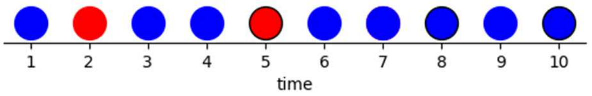

Fig. 1 Two malicious and eight standard signals over a 10-second time interval.

Let us consider a few exceptional cases to better understand the problem at hand. Suppose that $m = n$, i.e., the IDS checks the signal at every possible time step. Then, the probability of catching the malicious signal is 1. If $m = 0$, i.e., the IDS makes no checks, then the probability of detection is 0. In general, it stands to reason that as m increases, the probability of detection also increases. Similarly, as k increases, the probability of detection also increases. On the other hand, as n increases, the probability of detection decreases.

Since we assume that the signals are random and independent, the probability of detecting a malicious signal does not depend on whether the signals are sent at regular intervals or within a short time of each other. For example, suppose there are 20 equally spaced instances of a time interval at which signals can be sent. Suppose that an attacker sends three malicious signals. Then, the probability of detecting signals sent at $t = 5$, $t = 10$, and $t = 15$ would be the same as the probability of detecting signals sent at $t = 1$, $t = 2$, and $t = 6$. Thus, the probability of detecting a malicious signal does not depend on the pattern of the time instances at which the attacks take place. We now state our main result for calculating the detection rate by an IDS given a fixed number of attacks and checks.

Theorem 1. Let I be a time interval that is divided into n discrete time instances $\{t_i\}_{i=1}^{n}$ at which a malicious signal can occur. Let k be the number of malicious signals. Let m be the number of checks performed by IDS to detect a malicious attack. Then the probability of detecting x malicious signals is given by:

$$p(x) = \frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}} \tag{1}$$

Proof. We begin by counting the number of possible combinations of time instances for IDS checks. Note that due to the temporal nature of the instances, the IDS checks must occur in sequential, increasing order. Thus, every selection of m instances will correspond uniquely to a sequence of IDS checks. Since there are a total of n instances of which m are selected by IDS, the total number of combinations is given by $\binom{n}{m}$. Note that since the IDS checks are performed at random and independently, each combination of time instances is equally likely to occur.

Next, we count the combinations of instances at which the IDS catches x malicious attacks. Assume without loss of generality that a sequence of k attacks is fixed. Note that the sequence of time instances corresponding to a fixed pattern of an attack must be in increasing order. Thus, each combination of k malicious and $n - k$ regular signals correspond uniquely to an attack pattern. It follows that there are $\binom{k}{x}$ different ways the IDS can detect x attacks among a total of k. For each combination of x detected attacks there are $\binom{n-k}{m-x}$ combinations of regular signals tested by the IDS. We obtain that the total number of combinations of x malicious and m-x regular signals that can be tested by IDS is $\binom{k}{x}\binom{n-k}{m-x}$. Note that since the malicious signals are random and independent, each combination of time instances is equally likely to occur. It follows that the probability of detecting x malicious signals by the IDS is given by $\frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}}$. We can use Theorem 1 to calculate the overall probability of IDS detecting a malicious signal. The next corollary provides the result.

Corollary 2. In the same situation as in Theorem 1, the probability of IDS detecting at least one malicious signal is given by equation (2).

$$1 - p(0) = 1 - \frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}} \tag{2}$$

Proof. The result follows directly from Theorem 1.

To illustrate the utility of Corollary 2, consider the following example. Suppose that over the next one minute a hacker sends $k = 10$ malicious signals to a network. Assume that the signals can only be sent at 1-second intervals, i.e., $n = 60$. Suppose that an IDS will check for malicious signals at m randomly chosen instances. The graph in Fig. 2 shows the probability that the IDS will detect at least one malicious signal. The intrusion detection rate reaches nearly 100% once the number of checks exceeds 20. The simulation presented in Fig. 2 illustrates the benefits of the proposed probability framework. It shows that it is not necessary to check all $n = 60$ instances to catch an attack. An IDS can be designed to analyze only a fraction of the total network flow while achieving high detection rates. Thus, we obtain a faster and more efficient IDS.

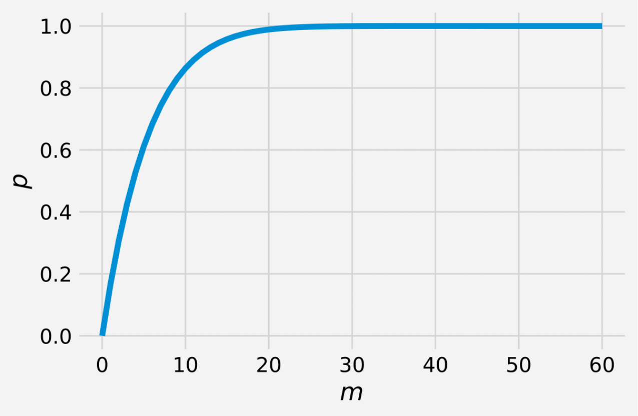

Fig. 2 The probability of detecting at least one malicious signal with m checks over n=60-time instances. The intrusion detection rate quickly rises to 100%.

Another important corollary of Theorem 1 is the expected number of detected malicious signals by an IDS. It helps our understanding of the average expected detection rate.

Theorem 3. In the same situation as in Theorem 1, the expected number of detected malicious signals by an IDS is given by the equation,

$$\mu = \frac{m}{n}.k \tag{3}$$

Proof. Let X be the number of attacks during the interval I. Then, by Theorem 1, the expected value of X is given by,

$$\begin{aligned} E[X] &= \sum_{x=0}^{k} x.p(x) \\ &= \sum_{x=0}^{k} x.\frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}} \\ &= \textstyle\sum_{x=0}^{k} x.\frac{k!}{(x-1)!(k-x)!}.\frac{\binom{n-k}{m-x}}{\binom{n}{m}} \end{aligned} \tag{4}$$

Next, we make substitutions y = x-1 and $l = k$-1. Then, Equation (4) can be rewritten and continued in the following manner,

$$\begin{aligned} &= \sum_{y=0}^{l} \frac{(l+1)k!}{y!\,(l-y)!}.\frac{\binom{n-l-1}{m-y-1}}{\binom{n}{m}} \\ &= \frac{l+1}{(n/m)}\sum_{y=0}^{l} \frac{l!}{y!\,(l-y)!}.\frac{\binom{n-l-1}{m-y-1}}{\frac{(n-1)!}{(m-1)!\,(n-m)!}} \end{aligned} \tag{5}$$

We make another pair of substitutions $s = n$-1 and $t = m$-1. Then, Equation (5) can be rewritten and continued in the following manner:

$$
\begin{aligned}
&= \frac{l+1}{(n/m)} \sum_{y=0}^{l} \frac{l!}{y!\,(l-y)!} \cdot \frac{\binom{s-l}{t-y}}{\frac{s!}{t!\,(s-t)!}} \\
&= \frac{km}{n} \sum_{y=0}^{l} \binom{l}{y} \frac{\binom{s-l}{t-y}}{\substack{s \\ t}} \qquad (6) \\
&= \frac{km}{n}
\end{aligned}
$$

The last step follows from the equality $\sum_{y=0}^{l} \binom{l}{y} \frac{\binom{s-l}{t-y}}{\substack{s \\ t}} = 1$ which is the statement about the total probability of all possible outcomes of a random experiment given in Theorem 1.

The result in Theorem 3 implies that the average expected number of intrusion detections is directly proportional to the number of checks and the number of attacks. As we increase the number of checks we expect to identify more attacks. Similarly, as the number of attacks increases the expected number of identified intrusions also increases. On the other hand, Theorem 3 implies that the average expected number of intrusion detections is inversely proportional to the total number of time instances at which a signal can be sent.

The conditions for detecting a malicious signal by an IDS can be modelled via the Poisson process. The Poisson point process is a counting process that represents the total number of occurrences or events that happen over a fixed period. The Poisson process is characterized by the Poisson distribution:

$$
p(x) = \frac{e^{-\mu} \mu^{x}}{x!}, \qquad (7)
$$

where μ is the average number of events during the fixed period and x is the number of events. In the context of IDS, the detected malicious signals over the fixed interval I represent the Poisson events. To apply the Poisson distribution, we need to determine the expected number of events which is given by Equation (3). Thus, the probability of detecting at least one malicious signal using the Poisson process is given by equation,

$$
1 - p(0) = 1 - e^{-\frac{mk}{n}} \qquad (8)
$$

The relationship between the approaches in Equation (2) and Equation (8) is illustrated in Fig. 3. As shown in the figure, the probability of attack modelled under the uniform process is close to that of Poisson process. We conclude that the Poisson process can be used effectively to approximate the true probability of detection.

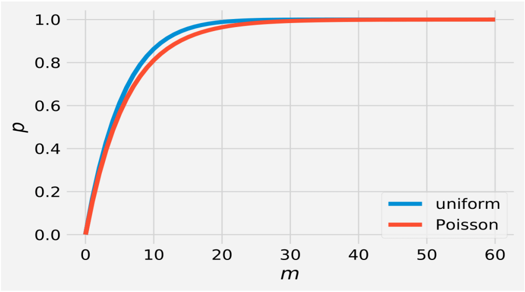

Fig. 3 Comparing the probability of detecting at least one attack under uniform and Poisson distributions.

III. Results and Discussion

In this section, we carry out a series of numerical experiments to verify the theoretical results obtained in Section 3. In particular, the experiments are designed to analyze the probability distribution of intrusion detection and the expected number of detected attacks. We begin by considering a situation where an interval of 1 minute is divided into n = 60 instances (seconds). A malicious signal can be sent to the network at any instance. Suppose that an attacker sends k = 5 malicious signals at different time instances. Assume that an IDS randomly checks m=10 instances to catch the attack. We simulate this scenario 1000 times and calculate the fraction of cases where exactly 2 malicious signals are detected by the IDS. We repeat the above experiment 10,000 times. The histogram of the resulting probabilities is presented in Fig. 4. For reference, we also calculate the theoretical probability p(2) using Equation 1. As shown in Fig. 4, the simulated detection probabilities are symmetrically distributed around the theoretical probability. Thus, the experiment results support the theoretical postulations developed in Equation 1.

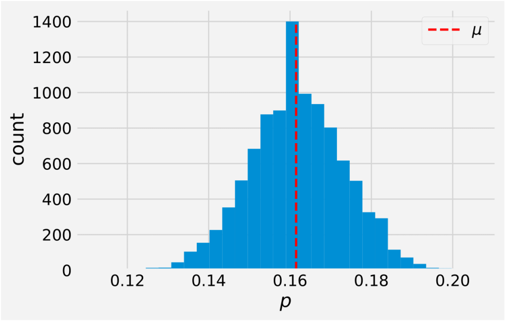

Fig. 4 A numerical simulation with a total of 10^7 runs. The simulation parameters are n=60, m=10, k=5, and x=2. The value μ=0.1615 is the theoretical probability obtained from Equation 1

Another aspect of IDS is the expected number of detected attacks. Assume the same scenario as above with n=60, m=10, and k = 5. Two ways to calculate the expected number of attacks exist: the direct approach and Equation 1. The direct calculation is done using the standard definition of the expected value:

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{k} x. \frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}} \\
&= \sum_{x}^{5} x. \frac{\binom{4}{x}\binom{55}{10-x}}{\binom{60}{10}} \qquad (9) \\
&= \frac{5}{6}
\end{aligned}
$$

A more efficient approach to calculate the expected number of attacks is given by us in Equation 1:

$$
E[X] = \frac{k\,.m}{n} = \frac{5\,.10}{60} = \frac{5}{6} \qquad (10)
$$

It follows from the above comparisons that the direct approach and the one provided by us in Equation 1 yield the same results.

To further test our theoretical results, we consider a scenario where a time interval is divided into n=100 instances.

An attacker can send a malicious signal at any instance. Suppose that the attacker decides to send k=20 malicious signals at different instances. Assume that the IDS is set to perform m=15 checks during the interval. Our goal is to calculate the probability that IDS detects x=4 malicious signals. We perform 10,000 experiments where each experiment consists of 1000 simulated runs.

The histogram of the resulting probabilities is presented in Fig. 5. For the reference, we calculate the theoretical probability p(4) using Equation 1. As shown in Fig. 5, the histogram is centered symmetrically around the theoretical probability.

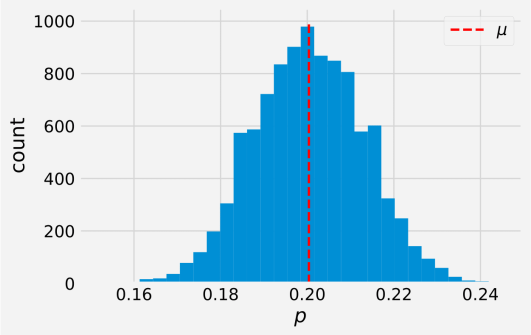

Fig. 5 A numerical simulation with a total of 10^7 runs. The parameters of the simulation are n=100, m=15, k=20, and x=4. The value $\mu = 0.2$ is the theoretical probability obtained from Equation 1

In addition, we calculated the expected number of detected attacks in the same scenario as above using the direct approach and Equation 3. Using the direct approach, we obtain:

$$\begin{aligned} E[X] &= \sum_{x=0}^{k} x.\frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}} \\ &= \sum_{x}^{20} x.\frac{\binom{20}{x}\binom{80}{15-x}}{\binom{100}{15}} \\ &= 3 \end{aligned} \tag{11}$$

A more efficient approach to calculate the expected number of attacks is given by us in Equation 1:

$$E[X] = \frac{k.m}{n} = \frac{20.15}{100} = 3 \tag{12}$$

It follows from the above comparisons that the direct approach and the one provided by us yield the same results. We observe that the experimental results support the theoretical findings from the previous section.

IV. Conclusion

The widespread adoption of network-based technologies has increased the potential for damage caused by a malicious attack on a network. Both the frequency and the severity of network attacks have risen over the past decade. As a result, it has become essential to develop effective intrusion detection systems. In this paper, we provide a probabilistic framework to analyze the detection rate of malicious attacks. We carried out careful theoretical and experimental analyses of the research problem. We developed the formula for calculating the intrusion detection rate for a fixed set of parameters. Given an interval of time that is divided into discrete instances at which the attacks can occur we can calculate the probability of IDS detecting x attacks via Equation 1. In addition, the expected number of detected attacks is also calculated via Equation 3. The theoretical results were tested and validated through numerical experiments. The outcome of the experiments confirmed the original theoretical results. We note that even with a simple strategy such as uniform sampling the probability of detecting at least one malicious attack is high given a small number of checks: the intrusion detection rate is nearly 0.9 when checking only 15% of time instances. The detection rate reaches 1 when checking 30% of time instances.

We believe that the probabilistic framework developed in this paper would be of use to IDS experts. For a network-based IDS that is checking the incoming packets in real time the computational cost of analyzing the entire traffic flow can be prohibitively expensive. So, checking only a probabilistically sampled fraction of the network traffic would allow the IDS to handle its task. However, checking only a fraction of the traffic introduces a probability of missing an attack. We hope to provide a better understanding of the likelihood of detecting an attack by an IDS and improve the design of the system. In practical terms, the results of this paper help IDS designers achieve appropriate detection rates while maintaining a low false alarm rate.

The groundwork that has been laid out in this paper can be used for future research into understanding of the probabilities related to intrusion detection. We believe that there are multiple avenues for future research that stem from the present work. The key assumption of our study is the uniform distribution of attacks and checks. However, it does not cover all the intrusion scenarios. It would be necessary to address other attack and detection patterns in future research.

References

[1] K. N. Sevis and E. Seker, “Cyber warfare: terms, issues, laws and controversies,” 2016 International Conference On Cyber Security And Protection Of Digital Services (Cyber Security), Jun. 2016, doi:10.1109/cybersecpods.2016.7502348.

[2] G. De Masi, “The impact of topology on Internet of Things: A multidisciplinary review,” 2018 Advances in Science and Engineering Technology International Conferences (ASET), Feb. 2018, doi:10.1109/icaset.2018.8376837.

[3] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, and N. Ghani, “Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet-Scale IoT Exploitations,” IEEE Communications Surveys &amp; Tutorials, vol. 21, no. 3, pp. 2702–2733, 2019, doi: 10.1109/comst.2019.2910750.

[4] Ventures C. Cybersecurity jobs report. Herjavec Group. 2017 May;1.

[5] A. Borkar, A. Donode, and A. Kumari, “A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS),” 2017 International Conference on Inventive Computing and Informatics (ICICI), Nov. 2017, doi:10.1109/icici.2017.8365277.

[6] P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, “Securing the Smart Grid: A Comprehensive Compilation of Intrusion Detection and Prevention Systems,” IEEE Access, vol. 7, pp. 46595–46620, 2019, doi: 10.1109/access.2019.2909807.

[7] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, “Intrusion detection system: A comprehensive review,” Journal of Network and Computer Applications, vol. 36, no. 1, pp. 16–24, Jan. 2013, doi:10.1016/j.jnca.2012.09.004.

[8] Y. Afek, A. Bremler-Barr, and S. L. Feibish, “Zero-Day Signature Extraction for High-Volume Attacks,” IEEE/ACM Transactions on Networking, vol. 27, no. 2, pp. 691–706, Apr. 2019, doi:10.1109/tnet.2019.2899124.

[9] R. Samrin and D. Vasumathi, “Review on anomaly based network intrusion detection system,” 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Dec. 2017, doi:10.1109/iceeccot.2017.8284655.

[10] S. Oshima, T. Nakashima, and Y. Nishikido, “Extraction for Characteristics of Anomaly Accessed IP Packets Based on Statistical Analysis,” Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007), Nov. 2007, doi: 10.1109/iihmsp.2007.4457652.

[11] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, “A Deep Learning Approach to Network Intrusion Detection,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 1, pp. 41–50, Feb. 2018, doi: 10.1109/tetci.2017.2772792.

[12] M. H. Ahmadzadegan, A. A. Khorshidvand, and M. G. Valian, “Low-rate false alarm intrustion detection system with genetic algorithm approach,” 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Nov. 2015, doi:10.1109/kbei.2015.7436188.

[13] S. Naseer et al., “Enhanced Network Anomaly Detection Based on Deep Neural Networks,” IEEE Access, vol. 6, pp. 48231–48246, 2018, doi: 10.1109/access.2018.2863036.

[14] F. Kamalov and F. Thabtah, “A Feature Selection Method Based on Ranked Vector Scores of Features for Classification,” Annals of Data Science, vol. 4, no. 4, pp. 483–502, Jul. 2017, doi: 10.1007/s40745-017-0116-1.. 6, pp. 48231–48246, 2018, doi:10.1109/access.2018.2863036.

[15] F. Kamalov, “Generalized feature similarity measure,” Annals of Mathematics and Artificial Intelligence, vol. 88, no. 9, pp. 987–1002, May 2020, doi: 10.1007/s10472-020-09700-8.

[16] F. Thabtah and F. Kamalov, “Phishing Detection: A Case Analysis on Classifiers with Rules Using Machine Learning,” Journal of Information &amp; Knowledge Management, vol. 16, no. 04, p. 1750034, Nov. 2017, doi: 10.1142/s0219649217500344.

[17] F. Kamalov and H. H. Leung, “Outlier Detection in High Dimensional Data,” Journal of Information & Knowledge Management, vol. 19, no. 01, p. 2040013, Mar. 2020, doi: 10.1142/s0219649220400134.

[18] A. Garg and P. Maheshwari, “A hybrid intrusion detection system: A review,” 2016 10th International Conference on Intelligent Systems and Control (ISCO), Jan. 2016, doi: 10.1109/isco.2016.7726909.

[19] C.-M. Ou, “Host-based Intrusion Detection Systems Inspired by Machine Learning of Agent-Based Artificial Immune Systems,” 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA), Jul. 2019, doi:10.1109/inista.2019.8778269.

[20] M. Ahmed, R. Pal, Md. M. Hossain, Md. A. N. Bikas, and Md. K. Hasan, “NIDS: A Network Based Approach to Intrusion Detection and Prevention,” 2009 International Association of Computer Science and Information Technology - Spring Conference, 2009, doi:10.1109/iacsit-sc.2009.96.

[21] Jianning Mai, A. Sridharan, Chen-Nee Chuah, Hui Zang, and Tao Ye, “Impact of Packet Sampling on Portscan Detection,” IEEE Journal on Selected Areas in Communications, vol. 24, no. 12, pp. 2285–2298, Dec. 2006, doi: 10.1109/jsac.2006.884027.

[22] Rong Cong, Jie Yang, and Gang Cheng, “Research of sampling method applied to traffic classification,” 2010 IEEE 12th International Conference on Communication Technology, Nov. 2010, doi:10.1109/icct.2010.5689208.

[23] J. M. C. Silva, P. Carvalho, and S. R. Lima, “Analysing traffic flows through sampling: A comparative study,” 2015 IEEE Symposium on Computers and Communication (ISCC), Jul. 2015, doi:10.1109/iscc.2015.7405538.

[24] I. Paredes-Oliva, P. Barlet-Ros, and J. Sole-Pareta, “Scan detection under sampling: a new perspective,” Computer, vol. 46, no. 4, pp. 38–44, Apr. 2013, doi: 10.1109/mc.2013.70.

[25] K. Bartos, M. Rehak, and V. Krmicek, “Optimizing flow sampling for network anomaly detection,” 2011 7th International Wireless Communications and Mobile Computing Conference, Jul. 2011, doi:10.1109/iwcmc.2011.5982728.

[26] G. Androulidakis, V. Chatzigiannakis, and S. Papavassiliou, “Network anomaly detection and classification via opportunistic sampling,” IEEE Network, vol. 23, no. 1, pp. 6–12, Jan. 2009, doi:10.1109/mnet.2009.4804318.

[27] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina, “Impact of packet sampling on anomaly detection metrics,” Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, Oct. 2006, doi: 10.1145/1177080.1177101.