On Accurate and Scalable Anomaly Detection in Next Generation Mobile Network

Fazirulhisyam Hashim and Abbas Jamalipour
The University of Sydney, NSW 2006, Australia
Email: {fhisyam,abbas}@ee.usyd.edu.au

Abstract—This paper proposes an adaptive sampling strategy to address the accuracy and scalability issues of anomaly detection at high-speed backbone side of Next Generation Mobile Network (NGMN). The proposed sampling strategy is formulated based on the network traffic condition. It is constituted by two important functions namely the traffic identification and the sampling decision. While the former utilizes spectral analysis to identify the severity status of the traffic flows, the latter exploits both the flow status and flow size to compute the optimal sampling rate. In addition, a renormalization process is proposed to address the scalability issue in the network. Our analysis demonstrates that the proposed technique is efficient in providing adequate statistics for detecting anomaly traffic and scales well to the high speed traffic of NGMN.

I. INTRODUCTION

Owing to the principle of providing seamless service continuity to its users, the Next Generation Mobile Network (NGMN) integrates various existing and future access networks via hierarchical architecture convention [1]. The complex architecture, however, exposes NGMN to external security threats that can be initiated from neighboring access networks. While the security threats can be mitigated through the cooperation of network entities at higher NGMN tiers (i.e., backbone network) [2], the enormous amount of disparity traffic at these tiers introduces scalability issue to the security solution. It is impractical for any network devices (i.e., network routers) at the high-speed backbone network to collect and store the information of every incoming packets. Although sampling is regarded as one of the best answers for effective scalability solution, it leads to inaccuracy issue in both quality of service (QoS) and security aspects, in particular on the traffic inference and anomaly detection, respectively. The inherent lossy characteristic of sampling (i.e., packets are discarded without inspection) has restricted its capability in providing a complete and reliable traffic statistic, thereby limiting the efficiency of both traffic inference and anomaly detection.

In recent years, an increasing amount of alternative traffic accounting methods has been proposed to address the QoS aspect of the network. Most of the literature have largely focused on analyzing traffic bias, devising better sampling techniques [3], and recovering statistics of the underlying traffic trace using inference [4][5]. In principal, a good traffic inference can be achieved by sampling the elephant flow (instead of mouse flow), which corresponds to a significantly large continuous flow that occupies a large portion of the total bandwidth over a period of time. While capturing the elephant flow improves the traffic inferences, it has significantly degraded the detectability of malicious anomaly due to their conflicting considerations. Note that malicious traffic is not necessarily appear in the same region with the traffic inference metrics, thus can also occur at mouse flow. Therefore, it is of paramount concern to address both traffic inference and anomaly detection without biasing towards either of them, as well as balancing the trade-off between the accuracy and overhead of measurements. The optimal sampling decision therefore should incorporate some forms of intelligence so that it can address the requirement of low false alarm, low computation complexity and resources optimization. While recent literature have shown the impact of sampling on anomaly detection metrics [6][7] and router resources (i.e., processor and memory) [8], none of them have considered both QoS and security aspects concurrently.

In this paper we propose an adaptive packet sampling strategy for accurate and scalable anomaly detection. The sampling strategy provides an optimal sampling decision by reacting appropriately to the network traffic behavior. It considers two important performance parameters namely the accuracy (i.e., reflects the traffic inference and anomaly detection) and efficiency (i.e., reflects the router resources). The proposed strategy utilizes spectral analysis to identify the behavior of the observed traffic which subsequently supplies two important information namely the severity level and the flow sizes. Upon receiving these information, the sampling decision is formulated to assign higher priority to the malicious elephant flow and lower priority to the normal mouse flow, thereby ensuring higher possibility of sampling the malicious dominant flows. Furthermore, the scalability issue is addressed by the proposed renormalization process in our sampling strategy. This process computes the number of cache entries (i.e., mainly from the least priority entries) that need to be freed in order to avoid cache overflow in the router memory. The formulation therefore improves the fairness and scalability issues by discarding the lower priority traffic in the cache. Thus, the router is still capable of sampling the incoming packets during heavy traffic (as is not the case for a conventional router).

The remainder of this paper is structured as follows. Section II briefly discusses the methodology of the proposed sampling strategy followed by a detailed description of its framework. Performance evaluation is discussed in
Section III, followed by some concluding remarks.

II. THE PROPOSED SAMPLING STRATEGY

The concept of the proposed sampling strategy is illustrated in Fig. 1. Generally, a two-step method of packet processing namely the traffic identification (i.e., implemented at the Training Phase as depicted in Fig. 2) and sampling decision (i.e., executed at the Triggering System) is applied to the incoming traffic. In the first step, the behavior of the observed traffic is identified as either normal or anomaly traffic. Meanwhile, the second step is responsible for determining the appropriate sampling rate. The identification method utilizes a flow-level analysis to investigate the spectral distribution of the incoming traffic and is carried out at every time interval. It involves exploiting the periodicity of the packet arrival process and deriving the attack signature by transforming the time-series data trace into the frequency domain. From the analysis, the status of the traffic behavior is reported in the form of attack probability (i.e., indicates possible presence of attack traffic). Furthermore, the process also provides the flow size information of the observed traffic. Consequently by using both information, the appropriate sampling decision (in the form of sampling rate) is formulated, and hence is used for the remaining period. Upon deciding on the appropriate sampling rate, the memory and processor react by determining the optimal number of entries need to be freed from the cache during the renormalization process.

In the following, the theoretical framework of the proposed sampling strategy is discussed in detail.

A. Traffic Identification

The traffic identification method stems from our previous work [9] on defining and classifying the traffic behavior based on its spectral characteristic using the Lomb periodogram [10]. The spectral analysis paradigm is capable of detecting possible attacks without requiring packet contents inspection, thus more efficient to be implemented at the core NGMN environment where delay performance constitutes one of the main parameters of interest. In addition, in [9] we have demonstrated the efficacy of spectral analysis to disclose the presence of malicious flows in aggregate traffic\(^1\). The distinct properties of spectral characteristic can be used to differentiate between the denial of service (DoS) and distributed DoS (DDoS) attacks. For DoS traffic, the dominant frequencies occupy higher frequency bands while for DDoS it occurs at lower frequencies. Interested readers may refer to [9] for a detailed description of the traffic classification mechanism. As mentioned earlier, the traffic identification process provides two important information for sampling decision namely the severity level and flow size of a particular flow. Let \(C_i\) be a counter dedicated to flow \(l\) (where \(l \subset m\)), and comprises of \(k\) numbers of subcounters \(c\) (as depicted in Fig. 3). For a specific duration of \(t_i\), the time line is divided into several subintervals such that \(C_i = \{c_j : j = 1, 2, \ldots k\}\). Given the scenario where \(A_i(j)\) represents the possible attack event (from traffic identification process) added to the specific \(j\)th subcounter of the flow \(l\), the severity of the flow (represented by the probability of the attack) can be expressed as \(Pr(A_i) = \frac{1}{\sum_{j=1}^{k} x_j}\) where \(x_j\) represents the detected \(A_i(j)\) event. The probability of the attack is bounded by the following condition, \(Pr(A_i) = \frac{1}{\sum_{j=1}^{k} x_j} \leq 1\). Since the attack probability constitutes one of the sampling decision parameter, the lower bound of the probability \((\frac{1}{k})\) is chosen so that the normal traffic is given a very low probability (instead of 0). Note that the normal traffic (i.e., corresponds to non-existence of the \(A_i(j)\) event in subcounters) will never be sampled if value 0 is used. Meanwhile, the number of subcounters \(k\) is a compromise between two opposing considerations. While a large value of \(k\) reduces the false alarm probability, a smaller value provides longer analysis period per bin, hence higher chances in capturing an attack. Due to dynamic nature of network traffic and topology, the number of the measurement bins should be kept as a configurable parameter.

\(^1\)Since the traffic is in aggregated form (i.e., containing malicious and non-malicious flows), filtering out the suspected malicious traffic at this stage may introduce significant number of false alarms to the system.
B. Sampling Decision

The sampling decision is formulated based on two important factors namely the accuracy and efficiency. While the former corresponds to the traffic inference and anomaly detection, the latter reflects the effect of sampling to router resources, in particular its memory.

1) Accuracy Parameter: The accuracy parameter indicates the preciseness of anomaly detection and traffic measurement subjects to a given sampling rate. While the anomaly detection is addressed by the severity level \( P_r(A_i) \) (as discussed in previous section), the traffic measurement is devised based on the proportion of a particular flow size over the entire flows population. Consider a set of \( m \) flows of various sizes \( S = \{ s_i : l = 1, 2, \ldots, m \} \), an elephant flow is classified if the proportion of its flow size \( \frac{s_l}{\sum s_i} \) is greater than a predefined threshold \[4\]. This ensures that the captured elephant flow is comprised by large packet and byte count flows as well as bursty flows. As mentioned earlier, for accurate anomaly detection the philosophy of the sampling decision should be governed by the level of severity and flow size parameters. Therefore, we formulate a mechanism in which any malicious elephant flow is given higher priority compared to any normal mouse flow. Let \( P_r(A_i) \) and \( s_l \) be the severity level and flow size of a particular flow \( l \), the prioritization can be computed as \( \chi_l \cdot \omega_l \) where \( \chi_l = P_r(A_i) \cdot s_l \) and \( \omega_l \) is the weight given to the \( \chi_l \). Since both the severity level and flow sizes are known parameters (i.e., from the traffic identification process), higher \( \omega_l \) is given to malicious elephant flows. Unfortunately due to the dynamic nature of network traffic, determining the exact \( \omega_l \) for a specific traffic condition can be an arduous task. Nevertheless, by exploiting the self-similarity attribute in network traffic [11], it is feasible to assume that the network operator should have a prior knowledge of the normal traffic profile. Hence, the appropriate \( \omega_l \) for a specific \( \chi_l \) can be predefined using a full histogram of the attack probability \( P_r(A_i) \) and flow size \( s_l \). However for simplicity and brevity purposes, in this paper we utilize the commonly used sampling rate (i.e., 0.1, 0.01, 0.02, etc.) in existing routers as the \( \omega_l \) parameter. Using a simple set of rules, a larger \( \omega_l \) value is assigned to a malicious elephant flow (i.e., large \( P_r(A_i) \) and \( s_l \) values), and conversely a normal mouse flow (i.e., small \( P_r(A_i) \) and \( s_l \) values) is given a relatively smaller value. Once the aforementioned steps are completed, the sampling rate for the remaining \( N[l_i] \) interval can be calculated as follows,

\[
\mathcal{R}[l_i] = \frac{\sum_{l=1}^{m} \chi_l \cdot \omega_l}{\sum_{l=1}^{m} s_l} \quad (1)
\]

The sampling rate is normalized by the the total number of flow sizes. This offers a trade-off between the opposing consideration in both anomaly detection and traffic measurement, i.e., appropriate sampling rate when massive attacks occur in a high traffic volume. From Eq. 1, it can be observed that the sampling decision is based on the overall traffic flows condition (i.e., due to packet sampling paradigm) with bias criteria towards accurate anomaly detection and traffic measurement.

2) Efficiency Parameter: The efficiency parameter reflects the competency of the sampling technique in addressing the scalability issue, thus implies the necessary response of the router memory in the presence of large number of incoming traffic (i.e., due to attacks or flash crowd events). It is essential to ensure that a certain number of entries are deleted so that the new entries created by incoming traffic do not push the size of the flow cache beyond the available memory. This scalability issue can be addressed by devising a renormalization process which is capable of dynamically adapt its cache entries to the sampling decision. Given the ratio between the new sampling rate (\( \mathcal{R}[l_i] \)) and the old sampling rate (\( \mathcal{R}[l_{i-1}] \)) as \( r = \frac{\mathcal{R}[l_i]}{\mathcal{R}[l_{i-1}]} \), the number of entries need to be freed is computed as follows,

\[
F = \begin{cases} 
\sum_{l=1}^{r=1/r} \frac{\eta_l (1 - rl)}{r} & \text{if } r < 1 \\
\sum_{l=1}^{r=1} \frac{\eta_l (1 - \frac{1}{rl})}{rl} & \text{if } r > 1 
\end{cases} \quad (2)
\]

where \( \eta_l \) is a pointer that represents the flow size of a particular \( \chi_l \) values inside the cache. By devising a data structure that sorts the \( \chi_l \) values in ascending order, the renormalization process ensures any normal mouse flow is given the least priority among all flows and therefore occupies a lower \( \eta_l \) value while the malicious elephant flow takes a higher \( \eta_l \) value. In the case of multiple flows exhibit similar value of \( \chi_l \), they will be grouped together into similar \( \eta_l \). This prioritization also ensures both the fairness and the accuracy parameter can be addressed in which the lower priority and less important flows will always be given a higher priority to be discarded.

III. PERFORMANCE EVALUATION

In this section we demonstrate the effectiveness of the proposed sampling strategy by determining the number of detected anomalies and the memory utilization at the router. Two independent real network traces (i.e., Trace A and Trace B, each trace consists of a 2-hour of traffic captured at OC-48 links) from CAIDA [12] are utilized as the normal traffic profile in our evaluation. The unsampled data sets are scaled to represent a high speed network (as is the case in NGMN). Eighteen random bandwidth attacks are initiated using the ns2 simulator and then are injected into both data sets.

Furthermore, due to the capability of offering low false positive rate, optimal detection accuracy and computing overhead, the Cumulative Sum (CUSUM) algorithm [13] is utilized in this work to capture the malicious traffic. In this algorithm any deviation from the mean traffic profile is considered as anomaly subjects to the condition \( y_n > Z \), where \( y_n \) and \( Z \) are the CUSUM test statistic parameter and threshold respectively.

A. The Impact of Static Sampling Technique

First, we examine the impact of the conventional static sampling rate (\( \mathcal{R}_{static} \)) to accuracy performance in particular the anomaly detection capability (as illustrated in Fig. 4). The performance is analyzed based on the success...
that for a lower sampling rate (i.e., 0.01 < \( \alpha \) ) it is worth to highlight that for a lower sampling rate (i.e., \( \alpha_{\text{static}} < 0.01 \)), the network suffers a significant degradation in accuracy performance (detection capability is less than 0.1), hence signifying the importance of having a suitable sampling rate for accurate anomaly detection.

B. Subcounters Parameter Tuning

Next, we empirically evaluate the appropriate number of subcounters \( k \) required in the counter \( C \) to sustain an acceptable detection accuracy. In conformance with the existing systems (e.g., Cisco NetFlow), in this paper we consider a 1-minute duration for the traffic identification \( t_i \). Fig. 5 depicts the variation of success detection ratio with the increase in the number of stages from 2 to 15. From the figure, the highest detection ratio can be achieved when \( k \leq 7 \). Nevertheless, it is worthwhile to highlight that the selection of the parameter \( k \) is influenced by the traffic behavior (particularly on the burstiness attributes), Lomb periodogram capability and also the duration of the traffic identification process \( t_i \). Since the challenging 1-minute is used as the duration for \( t_i \), we operate with \( k = 5 \) (i.e., 12 seconds for each subcounter) for the rest of the paper for our evaluations.

C. The Impact on the Accuracy Parameter

Fig. 6(a) and Fig. 6(b) depict the unsampled traffic for both Trace A and Trace B respectively, while their counterparts the sampled traffic (i.e., via adaptive sampling strategy) are illustrated in Fig. 6(c) and Fig. 6(d).

Note that the vertical dotted lines represent the detected anomalies and the gray shaded bars imply the undetected anomalies in the network. Interestingly, from the figure it can be observed that the proposed sampling technique is incapable of detecting all the 18 malicious anomalies, with only 16 (0.89 success ratio) and 17 (0.94 success ratio) in Trace A and Trace B, respectively. From our discrete observation, we found that the false negatives (at \( t = 101 \) sec and 230 sec for Trace A and \( t = 4145 \) sec for Trace B) are caused by the limitation of the predefined threshold value, \( Z \) in CUSUM algorithm (refer to Fig. 7). The undetected anomalies in both data sets are successfully sampled by the proposed sampling strategy, nonetheless the number of accumulated malicious packets is less than the threshold value. This finding therefore suggests the importance of investigating the correlation between a particular anomaly detection mechanism and sampling parameter (i.e., a specific sampling strategy will have distinct impact on different anomaly detection methods). Nevertheless, despite possessing lower accuracy performance to the one offered by \( \alpha_{\text{static}} = 0.1 \) (as indicated in Fig. 4), the adaptive sampling technique is preferred due to its effective scalability solution (to be discussed in the next section).

D. The Impact on the Efficiency Parameter

Given that higher sampling rate constitutes better accuracy performance (as verified in Sec. III-A), for comparison purposes we only present the memory utilization of \( \alpha_{\text{static}} = 0.1 \). Note that from our studies (the result is omitted due to space limitation) we found that lower sampling rate leads to less number of cache overflows, thereby improving its memory utilization. Fig. 8(a) and Fig. 8(b) illustrate the cache entry status inside the router memory of \( \alpha_{\text{static}} = 0.1 \) for Trace A and Trace B, respectively. It can
be observed that the memory utilization of $R_{static} = 0.1$ suffers severe degradation of efficiency performance and experiences large number of cache overflow (indicated by the black dotted marks). Meanwhile the cache status of adaptive sampling for Trace A and Trace B are given by Fig. 8(c) and Fig. 8(d), respectively. From the figures, in comparison to $R_{static} = 0.1$, it is apparent that the proposed sampling decision is capable of avoiding the cache overflow, thereby validating the efficiency of the technique.

E. Sampling Rate Distribution

Furthermore we also analyze the sampling rate distribution of the proposed technique. The idea is to examine the response of the sampling decision over the traffic behavior. Fig. 9 illustrates the distribution of the adaptive sampling rate throughout the simulation process. From the figure it can be observed that both traces possess significantly high sampling rate with average rate of Trace A is 0.102 (1 out of 10 packets) and Trace B is 0.125 (1 out of 8 packets). Note that due to the existence of many elephant flows and considerably large amount of attacks throughout the simulation, the sampling rate distribution is concentrated at high sampling rate, and thus verify the philosophy of the proposed technique.

IV. CONCLUSION

We have proposed an adaptive sampling strategy that is capable of detecting malicious traffic and efficiently utilizing router resources. Three important factors namely the anomaly detection, traffic measurement, and router resources have been considered in formulating the proposed sampling strategy. Two performance parameters namely the accuracy and efficiency have been defined to represent the aforementioned factors. The preliminary simulation results verify the efficacy of the proposed sampling strategy for both performance parameters, thereby validating our claim about the effectiveness of the proposed sampling strategy over the conventional technique.

REFERENCES