Journal of Computer and System Sciences ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Reducing false positives of network anomaly detection by local adaptive multivariate smoothing

Martin Grill [a,b,*], Tomáš Pevný [a,b], Martin Rehak [a,b]

[a] Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
[b] Cisco Systems, Inc., United States

ARTICLE INFO

Article history:
Received 31 August 2015
Received in revised form 6 March 2016
Accepted 17 March 2016
Available online xxxx

Keywords:
Network anomaly detection
Regression smoothing
False positive rate reduction

ABSTRACT

Network intrusion detection systems based on the anomaly detection paradigm have high false alarm rate making them difficult to use. To address this weakness, we propose to smooth the outputs of anomaly detectors by online Local Adaptive Multivariate Smoothing (LAMS). LAMS can reduce a large portion of false positives introduced by the anomaly detection by replacing the anomaly detector's output on a network event with an aggregate of its output on all similar network events observed previously. The arguments are supported by extensive experimental evaluation involving several anomaly detectors in two domains: NetFlow and proxy logs. Finally, we show how the proposed solution can be efficiently implemented to process large streams of non-stationary data.

1. Introduction

Increasing number and sophistication of attacks against critical enterprise computing infrastructure drives the need to deploy increasingly more sophisticated defense solutions. An essential component of the defense is Intrusion Detection System (IDS) [1] analyzing network traffic that crosses the defense perimeter and looking for evidence of ongoing malicious activities (network attacks). When such activity is detected, an alarm is raised and then analyzed by network administrator who determines the existence and scope of the damage, repairs it and improves the defense infrastructure against future attacks.

IDS can be categorized according to different criteria: by the location (on a host, a wired network, or a wireless network), detection methodology (signature matching, anomaly detection, or stateful protocol analysis), or capability (simple detection or active attack prevention) [1]. In context of our work, the most important criterion is the categorization by detection methodology, which is described in more detail below.

Signature matching techniques identify attacks by matching packet contents against specific attack signatures. The signatures are created using already identified and well-described attack samples, which are time consuming and can take from couple of hours up to several days, which gives attackers plenty of time for their criminal activities. The biggest weakness of this solution is that it detects only known attacks, which can be due to smart evasion techniques used by malware limiting. With the growing proportion of encrypted traffic, use of self-modifying malware and other evasion techniques, the use of detection techniques tailored to catch predefined known set of attacks is becoming increasingly irrelevant.

* Corresponding author at: Czech Technical University in Prague, Faculty of Electrical Engineering, Zikova 1903/4, 166 36 Prague 6, Czech Republic.
E-mail addresses: magrill@cisco.com (M. Grill), tpevny@cisco.com (T. Pevný), marrehak@cisco.com (M. Rehak).

http://dx.doi.org/10.1016/j.jcss.2016.03.007

Anomaly-based detection tries to decrease the human work (e.g. manual creation of signatures) by building statistical model of a normal behavior and detect all deviations from it. This enables to detect new, previously unknown attacks provided that their statistical behavior is different from that of the normal traffic. While anomaly-based methods are attractive conceptually, they have not been widely adopted. This is because they typically suffer from high false alarm rate (not every anomaly is related to the attack) rendering them useless in practice, since network operator can analyze only few incidents per day [2,3]. Decreasing the false positive rate is therefore important to make anomaly-based IDS competitive with signature matching based solutions.

Rehak [4] divides false positives of anomaly-based IDS into two classes: *unstructured* false positives are essentially a random noise caused by the stochasticity of the network traffic and *structured* false positives caused by a persistent but a very different behavior of a small number of network hosts, for example mail or DNS servers (the precise definition is left to Section 2).

This work proposes a method designed to decrease the rate of unstructured false positives by smoothing anomaly values with respect to time. This causes similar anomalies[1] occurring at different times to receive similar anomaly score, even though one would be otherwise flagged as anomaly. The rate of structured false positives is either decreased or remains the same, which depends on circumstances discussed in detail in Section 3. The method is evaluated with two different anomaly detection based IDSs, in both cases improving their accuracy.

The contribution of the paper is threefold.

- First, it mathematically formulates structured and unstructured false positives and argues why unstructured false positives are more difficult to white-list or remove.
- Second, it proposes a theoretically sound method to decrease the rate of unstructured false positives.
- Third, it shows how the method can be implemented to efficiently process data-streams.

This paper is organized as follows. The next section first properly defines classes of false positives, describe the proposed solution and mathematically proofs its correctness under mild assumptions. The same section also discusses the modification to efficiently process data-streams. Section 4 shows, how the method was deployed in two Intrusion detection systems. The related work is discussed in Section 5 while Section 6 concludes the paper.

2. Classification of false positives

Hereafter it is assumed that the network anomaly detection system (AD) observes a stream of network events (e.g., NetFlow [5], HTTP connections, etc.) produced by a set of hosts within the network. Anomaly detection system maintains internal model(s) to assign a score in range $[0, 1]$ to each observed event with zero indicating the normal event and one indicating possible attack, as it is assumed that malicious activities have statistical characteristics different from the normal ones [6–8] making them rare. As mentioned in the introduction the anomaly detection systems produce false alarms since an overwhelming majority of rare events are not caused by any attack. Rehak [4] divides these false positives into following two classes:

- *Unstructured false positives* are short-term events distributed uniformly over all the network hosts proportionally to the traffic volume. They are typically triggered by widespread, uniformly distributed behaviors (such as web browsing) and we model them as white noise (zero mean and finite variance) added to the anomaly detector's output. Therefore, the observed anomaly score y_i of an event x_i is equal to

$$y_i = g(x_i) + \eta_i, \tag{1}$$

  where $g(x_i)$ is the true anomaly score on event x_i and η_i is the additive white noise. The η_i therefore hides the true value $g(x_i)$.
- *Structured false positives* are caused by a (long-term) legitimate behaviors of a small number of network hosts. These behaviors are different from the background, and because they are found only at a very small portion of network hosts, they are reported as anomalies. Examples are rare applications performing software update, regular calls of unusual network APIs, etc. Since this type of false positive is typically limited to a small number of network hosts and its behavior is very regular, it can be quickly identified and filtered out using white-lists. Nevertheless, these white-lists are specific for a given network and are difficult to create before deployment. We define the structured false positives to be generated by a mixture of distributions

$$x_{sfp} \sim \frac{\sum_{j=1}^{m} \beta_j \epsilon_j(x_i)}{\sum_{j=1}^{m} \beta_j}, \tag{2}$$

  where ϵ_j represent structured false positive with weight β_j. Each component ϵ_j has small variance comparing to that of the unstructured false positives, but means of the components are typically far from each other.

[1] Similarity can be an arbitrary function $k : \mathcal{X} \times \mathcal{X} \mapsto [0, 1]$.