Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

A training-resistant anomaly detection system

Steve Muller [a,*], *Jean Lancrenon* [a], *Carlo Harpes* [a], *Yves Le Traon* [b], *Sylvain Gombault* [c], *Jean-Marie Bonnin* [c]

[a] *itrust consulting s.à r.l., Niederanven, Luxembourg*
[b] *University of Luxembourg, Luxembourg, Luxembourg*
[c] *IMT Atlantique, IRISA, UBL, Rennes, Bretagne, France*

ARTICLE INFO

Article history:
Received 24 October 2017
Received in revised form 16 January 2018
Accepted 23 February 2018
Available online 6 March 2018

Keywords:
Training attack
Intrusion detection system
Anomaly detection
Network security
Machine learning

ABSTRACT

Modern network intrusion detection systems rely on machine learning techniques to detect traffic anomalies and thus intruders. However, the ability to learn the network behaviour in real-time comes at a cost: malicious software can interfere with the learning process, and teach the intrusion detection system to accept dangerous traffic. This paper presents an intrusion detection system (IDS) that is able to detect common network attacks including but not limited to, denial-of-service, bot nets, intrusions, and network scans. With the help of the proposed example IDS, we show to what extent the training attack (and more sophisticated variants of it) has an impact on machine learning based detection schemes, and how it can be detected.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Intrusion detection is a quite old research topic (the first papers being published in the 1980's (Anderson, 1980; Denning, 1987)), yet it still constitutes an actively researched domain of computer security, especially in the field of cyber-physical systems such as Supervisory Control and Data Acquisition (SCADA) systems or Advanced Metering Infrastructures (AMI) (Mitchell and Chen, 2014). Over the past few years, the increasing interest in machine learning techniques led to the development of more sophisticated, so-called *anomaly* detection systems, which learn the "typical" behaviour of a monitored network or system. That way, they are able to spot deviations from the normal behaviour and thus, to a certain extent, detect previously unseen attacks.

Given the fact that a lot of different IDS strategies have been proposed over the years (Axelsson, 2000; Milenkoski et al., 2015), it is important to choose the one that really suits the needs. For instance, anomaly detection systems typically require the monitored network to be sufficiently static and predictable. While this is not necessarily the case for arbitrary computer networks, cyber-physical systems usually *do* meet this requirement, so a lot of research (Mitchell and Chen, 2014) has been conducted over the past few years in developing and improving on intrusion detection techniques for cyber-physical systems (Zhu and Sastry, 2010).

In addition, an automated learning system commonly requires a supervised initial training phase, during which it is faced with (manually labelled) benign and malicious data so that it learns the difference between these two data sets. Naturally, for optimal results, the learning process should be carried

* *Corresponding author.*
E-mail address: steve.muller@itrust.lu (S. Muller).
https://doi.org/10.1016/j.cose.2018.02.015
0167-4048/© 2018 Elsevier Ltd. All rights reserved.

out directly in the target network, and not in a lab. Nevertheless, many researchers use recorded data sets (such as the KDD'99 (KDD Cup, 1999) data set) to evaluate the performance of their anomaly detection algorithm. Unfortunately, the latter data sets are too generic to be actually used to train and deploy an intrusion detection system in a real network. This common practise can be explained by the fact that the used protocols are often proprietary or unknown, and that the network infrastructure is too complex, undocumented, or not available as a testing environment.

Moreover, an ideal intrusion detection system would spot undesirable content without requiring a training phase, since it can then be directly deployed in any production environment that is not known beforehand. In the machine learning domain, some schemes already exist which autonomously tell "normal" data apart from outliers, and which are thus suitable for intrusion detection (Buczak and Guven, 2016). For our purposes, clustering-based schemes seem to be the most promising ones, since they are unsupervised, relatively light-weight from a computation point of view (which is important if one wishes to build a real-time intrusion detection system), and allow multiple behaviours to be modelled at the same time (in contrast to Bayesian statistics, which merely splits the data into "normal" and "abnormal"). Moreover, they yield comprehensible results, in contrast to e.g. neural networks, where it is not so clear *why* they gave a certain output.

For machine learning based intrusion detection techniques, a lot of research has been made over the years, that increased their performance, their reliability, and their scope. However, attacks are also becoming more and more sophisticated. The most developed of them are referred to as advanced persistent threats (APTs): they cover all kinds of hacking or spying activities that are particularly stealthy and persistent (Cole, 2012). Given the fact that most networks and computer systems rely on anti-virus agents and intrusion detection systems, a lot of money and effort are put now into evading these security mechanisms (Cole, 2012). Automated learning systems (and especially those that continuously adapt to live data) are particularly affected by this fact, because their learning process can often be manipulated in such a way to make them progressively used to malicious data. This process is referred to as the *training attack*.

It is virtually impossible to design an intrusion detection system that defends against all modes of operation of APTs (and this is especially true when they are targeted, and thus human-operated). Therefore, in order to better understand how stealthy and long-term attacks act on a computer network, this paper focuses on a concrete example of an evasion technique that may be used by an advanced persistent threat, namely the training attack. To the best of our knowledge, very little research has been done to date, that analyses the robustness of intrusion detection systems against such evasion techniques.

The rest of this paper is organised as follows. Section 2 describes related work. In the following Sections 3 and 4 the threat model and the detection strategies are discussed, respectively. In particular, our proposed IDS is outlined in Section 4.3. The importance of the right parameter values is addressed in Section 5, and their choice is evaluated in Section 6. The paper concludes with Section 7.

2. Related work

Several other authors have adopted the approach of applying clustering techniques to a data stream collected from network (meta) data. A related area of research is the realm of stream clustering algorithms, which are able to cluster stream data (such as network metadata) on-the-fly. A comprehensive state-of-the-art has been recently composed by Ghesmoune et al. (2016), and is out of the scope for this paper. It is, however, not obvious how the resulting clusters can be interpreted in terms of intrusions or anomalies, and the related research area is also comparatively young.

For example, Tomlin et al. (2016) propose an IDS based on k-means and fuzzy cognitive maps (FCMs) that is applied to security events of a power system. They manage to improve on the detection accuracy of existing clustering-based IDS, but they still require a training data set for initial learning.

Hendry and Yang (2008) do not design an intrusion detection system per se, but they introduce an algorithm that uses data clustering to create attack signatures from recorded data. Unfortunately the algorithm needs to pre-process the data so that it cannot be used for on-line detection.

Similarly, Leung and Leckie (2005) develop a density-based clustering algorithm (called fpMAFIA), but it also requires a supervised learning session.

In contrast, Zhong et al. (2005) use an on-line variant of the k-means algorithm to group metadata of WLAN traffic into k clusters, for a fixed k. Any data point that is too far away from the center of the largest cluster, is considered an anomaly. While this approach is completely unsupervised, it has comparatively low detection rates of 65%–82%.

A similar approach is adopted by Alseiari and Aung (2015), who use a simplified k-means variant ("mini-batch k-means") to split smart meter readings into clusters; if a cluster is smaller than a given fixed value, it is considered anomalous. Most interestingly, they also account for the clusters' evolution by applying a sliding time window to the data. The authors claim to get slightly better detection rates, but also admit that the reported results are sometimes unreliable (100% false positive rate) and that more research is needed to tackle the issue.

An emerging research topic deals with the problem of training attacks that try to fool the intrusion detection systems by progressively manipulating the data they monitor. Wagner and Soto (2002) were the first to raise awareness about the issue; Barreno et al. (2006) explored the topic in more detail for intrusion detection systems.

Some authors focused on the related *mimicry attack*, which consists in evading the IDS, but not manipulating it permanently. Among them, Stevanovic and Vlajic (2014) disseminate on real-world occurrences of mimicry attacks for the case of anti-DDoS systems. The solution they propose consists in applying two independent anomaly detection systems: one is classical for DDoS detection, and one which particularly focuses on mimicry attacks.

Yen and Reiter (2008) describe an IDS which detects certain "stealthy" malware by monitoring the similarity in behaviour

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات