Summer Internship Report

Towards Reliable Evaluation of Anomaly-based Intrusion Detection Performance

Arun Viswanathan
USC/Information Sciences Institute
Email: aviswana@isi.edu

Mentors

Dr. Kymie Tan Bryan Johnson
Jet Propulsion Laboratory, California Institute of Technology

August 20, 2012

Abstract

This report describes the results of research into the effects of environment-induced noise on the evaluation process for anomaly detectors in the cyber security domain. This research was conducted during a 10-week summer internship program from the 19th of August, 2012 to the 23rd of August, 2012 at the Jet Propulsion Laboratory in Pasadena, California. The research performed lies within the larger context of the Los Angeles Department of Water and Power (LADWP) Smart Grid cyber security project, a Department of Energy (DoE) funded effort involving the Jet Propulsion Laboratory, California Institute of Technology and the University of Southern California/Information Sciences Institute. The results of the present effort constitute an important contribution towards building more rigorous evaluation paradigms for anomaly-based intrusion detectors in complex cyber physical systems such as the Smart Grid.

Anomaly detection is a key strategy for cyber intrusion detection and operates by identifying deviations from profiles of nominal behavior and are thus conceptually appealing for detecting "novel" attacks. Evaluating the performance of such a detector requires assessing: (a) how well it captures the model of nominal behavior, and (b) how well it detects attacks (deviations from normality). Current evaluation methods produce results that give insufficient insight into the operation of a detector, inevitably resulting in a significantly poor characterization of a detectors performance. In this work, we first describe a preliminary taxonomy of key evaluation constructs that are necessary for establishing rigor in the evaluation regime of an anomaly detector. We then focus on clarifying the impact of the operational environment on the manifestation of attacks in monitored data. We show how dynamic and evolving environments can introduce high variability into the data stream perturbing detector performance. Prior research has focused on understanding the impact of this variability in training data for anomaly detectors, but has ignored variability in the attack signal that will necessarily affect the evaluation results for such detectors. We posit that current evaluation strategies implicitly assume that attacks always manifest in a stable manner; we show that this assumption is wrong. We describe a simple experiment to demonstrate the effects of environmental noise on the manifestation of attacks in data and introduce the notion of attack manifestation stability. Finally, we argue that conclusions about detector performance will be unreliable and incomplete if the stability of attack manifestation is not accounted for in the evaluation strategy.

1 Introduction

The novelty and complexity of security threats to cyber systems, and more recently to cyber- physical systems, is growing at an alarming rate. Recent incidents such as the cyber attacks by foreign-military personnel on US Landsat and Terra spacecraft [1] and the shutdown of nuclear reactor centrifuges by the Stuxnet malware [4] highlight the importance of intrusion detection techniques that can detect anomalous adversarial behavior and zero-day attacks. A promising approach for detecting unseen threats that has received wide attention in the computer security domain is anomaly detection. An anomaly detector works in two main phases: a *learning phase* and a *detection phase*. In the learning phase, the detector learns the nominal behavior of a system by observing data representing normal or "non-malicious" system activity, while in the detection phase, the detector applies its learnt model of nominal behavior over subsequent data streams to report any deviations as anomalies or attacks. Although conceptually appealing for detecting zero day attacks, the lack of rigorous and reliable evaluation strategies for assessing anomaly detector performance has posed a great challenge with respect to its adoption into real-world operating environments [6, 8]. In this research, we explore the requirements necessary for a well-grounded evaluation of anomaly detectors performance and then focus on understanding a small, but important, subset of the evaluation challenge, namely, the effects of environment-induced variability on attack signals and their impact on the quality of assessment results for anomaly-based intrusion detectors.

The typical procedure for evaluating an anomaly detector consists of first training an anomaly detector over some representative data from a system and then evaluating its performance by testing its detection capabilities against different types of attacks and normal activity. Detector performance is then measured using *hit* and *miss* scores based on the number of attacks detected and the number of attacks missed. Unfortunately, this assessment is often unreliable and unrepresentative of the actual performance of a detector. For example, hit or miss scores do not convey why a detector performed badly or suggest any reasons for why the performance would remain the same in a different operating environment. As we discuss later in Section 2, a comprehensive evaluation regime must consider many different constructs for a reliable assessment.

In this work, we discuss only the environment-related constructs of evaluation. Operating environments can introduce variability in the data stream that can affect the learning and detection phases of a detector. Prior research has focused on understanding the impact of this variability in training data for anomaly detectors and its subsequent impact on detector performance. For instance, the work of Lee et al. [7] focused on applying information-theoretic measures to characterize the regularity of audit data, where regularity is defined with respect to the redundancies and sequential dependencies present in the data. Such characterization was then used to suggest improvements to existing anomaly detection models, to explain why existing models work and explain the performance of those models.

Unfortunately, there is no reported work on understanding the effects of the operational environment on the capabilities of a detector. Current evaluation approaches assume that if a detector does not detect an attack, then it is a miss as far as the evaluation is concerned. We argue that this is not necessarily true. Our hypothesis is that it is possible for the environment to introduce artifacts into the attack signal making the attack appear normal to the detector. Thus, if this is true, the causal

root of a miss assigned to the detector may not be detector failure but rather the result of the operating environment in which the detector was deployed. We argue that a reliable characterization of a detector's performance cannot be made if attacks do not manifest stably in the data stream.

Our objective here is to first prove or disprove the hypothesis that attack signals can manifest unstably due to environmental noise. To this end, we designed a simple controlled experiment, described in Section 3.1, to understand the different ways in which environmental noise can impact the manifestation of attack in the simplest of operating environments. Our preliminary findings are encouraging and are discussed in Section 3.3.

There are two main contributions of this research: (a) a preliminary taxonomy of important evaluation constructs for reliably evaluating an anomaly detector, and (b) an initial experimental proof of the hypothesis that attacks can manifest unstably in data due to environment-induced noise. Further, we also establish the need for understanding the stability of attack manifestations in data for accurately evaluating an anomaly detector's performance. The overall contribution of this work lies in improving the state-of-the-art in rigorous evaluation paradigms for cyber intrusion detectors by introducing a critical, but thus far missing, factor to the evaluation process – the stability of the attack signal.

2 Reliable Evaluation of an Anomaly Detector

The purpose of an evaluation is to gain insight into the workings of a detector. Specifically, as Paxson et al. [8] point out, a sound evaluation should answer the following questions: *(a) What can an anomaly detector detect?*, *(b) Why can it detect?*, *(c) What can it not detect? Why not?*, *(d) How reliably does it operate?* and *(e) Where does it break?*. As already discussed in Section 1 current evaluation approaches are inadequate to answer the above questions. For instance, using current approaches to evaluation, it is difficult to understand the reasons why a detector missed an attack. It is not clear if the miss should be genuinely attributed to the detector or to environmental causes that caused the attack to appear normal to the detector.

Figure 1 presents our preliminary taxonomy of evaluation constructs for evaluating detector performance. At a high-level, we see that the performance of a detector is dependent on the performance of both its learning and detection phases. We further discuss constructs related to each of those phases separately.

2.1 Evaluation of learning phase

The performance of the learning phase depends on the training data and the learning algorithm employed.

2.1.1 Training Data

The training data is used by a detector to learn the nominal behavior of a system. There are atleast three different aspects of the training data that can affect performance.

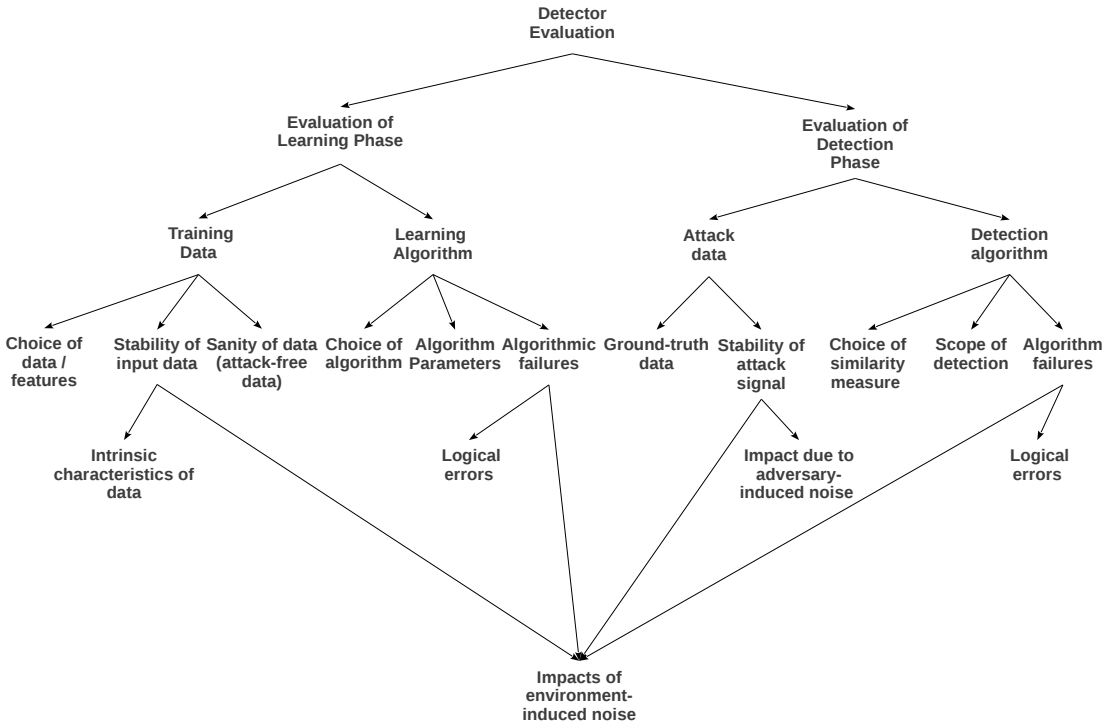

Figure 1: Different constructs for reliable evaluation of anomaly detector performance.

Choice of data / choice of features

An anomaly detector can detect attacks over multiple types of data and over different features of the data. But, certain data types and certain features within that data are more suitable for distinguishing normal behavior from anomalies and the choice indirectly affects the overall performance of a detector. This aspect has been well understood and has been researched thoroughly in the literature [2].

Stability of training data

As discussed by Lee et al. [7] and Paxson et al. [8], the basic premise for anomaly detection rests on an assumption that there exists some stability or regularity in training data that is consistent with the normal behavior and thus distinct from the abnormal behavior. A highly variable training data causes the detector to mis-learn the nominal behavior and which necessarily affects its performance.

Further, the stability of training data is influenced either by (a) the intrinsic characteristic of data or, (b) the operating environment-induced noise. For example, if the training data has attributes that are intrinsically random, say TCP sequence numbers, then that data is highly unstable. On the other hand, the data itself may be stable but might be affected by variability introduced by the operating environment. For example, some applications

sporadically generate spurious network traffic which could change the nominal behavior of a system as perceived by a detector.

Sanity of training data

A reliable evaluation of the learning phase must ensure that the detector is trained with attack-free data [8, 6]. If the training data is polluted with attacks or anomalies, the detector learns the anomalies as nominal behavior thus preventing it from detecting those anomalies.

2.1.2 Learning Algorithms

There are atleast three different constructs of the learning algorithm that can affect performance.

Choice of algorithm

Certain learning algorithms are better suited for detecting certain anomalies than others and thus influence the performance [3].

Algorithm parameters

The learning algorithms are heavily influenced by their parameters. For example, in the seminal work by Forrest et al. [5], the value of *window size* parameter was a deciding factor for the performance of the anomaly detector.

Algorithmic failures

The performance of a detector can be impacted by algorithmic failures which could either be failures due to logical errors in the algorithm implementation or errors that might be induced by the environment. We have not seen this aspect being evaluated explicitly in prior research but we believe this is an important factor to consider especially since different algorithms react differently to noise.

2.2 Evaluation of detection phase

The performance of the detection phase depends on attack data fed to the detector along with the detection algorithm employed.

2.2.1 Attack Data

The attack signal is a sequence of anomalous events and is input to the detector to evaluate the detection capabilities of a detector. There are two main constructs of attack data that affect performance of a detector.

Ground truth

The ground truth data consists of data clearly labeled as attack or normal data and is crucial for a reliable evaluation of a detector. It is often the case that such crisply labeled data is not available for evaluation in the computer security domain.

Stability of attack signal

A stable attack signal is one which appears sufficiently distinct from normal behavior to a detector under all circumstances. An attack signal could manifest unstably due to two main reasons.

- **Artificial adversary-induced noise**, wherein an attacker might distort an attack signal by generating artificial noise that makes the attack signal appear normal to a detector. Wagner et al. [9] have previously shown how *mimicry attacks* can confuse a detector.
- **Operating environment-induced noise**, wherein an attack signal might get distorted because of the variability introduced by the operating environment. This category of noise has been largely ignored and we discuss it further in Section 3. We also demonstrate that this is an important aspect for reliable evaluation.

2.2.2 Detection Algorithm

The attack signal is a sequence of anomalous events and is input to the detector to evaluate the detection capabilities of a detector. There are two main constructs of attack data that affect performance of a detector.

Choice of similarity measure

Anomaly detectors use a similarity measure to detect deviations of behavior from a learnt behavior profile. It is well-understood that the choice of the similarity measure greatly influences the performance of a detector [2].

Scope of detection

Every anomaly detector is limited to detect only a few types of anomalies or attacks. An understanding of this boundary is essential to reliably evaluate a detector.

Algorithmic failures

Similar to the learning phase, the performance of a detector can also be impacted by algorithmic failures in detection. Again, we have not seen this aspect being evaluated explicitly in prior research.

2.3 Observations on current state-of-the-art

We make three observations on the current state-of-the-art based on the evaluation taxonomy presented in Figure 1.

- We observe that a lot of existing evaluation efforts have focused on evaluating the learning phase of a detector and have ignored the detection phase of a detector.
- Even within the learning phase, most evaluation constructs have been studied in isolation of others. For instance, an evaluation that focuses on stability of training data does not necessarily evaluate the algorithmic constructs in the learning phase.

- We see that the operating environment significantly influences factors across both the learning and detection phases. Current methods do not address this adequately, but, we believe that an understanding of the operating environment is critical to the assessment of a detector.

3 Stability of Attack Manifestations

In this section, we describe our terminology and experimental methodology for proving our hypothesis on attack manifestations, introduce the notion of attack manifestation stability, describe its relationship to noise, and briefly discuss our preliminary results.

3.1 Experimental Methodology

The objective of the experiment is to observe the manifestation of an attack in symbolic data in the presence of noise introduced by the environment. In this experiment, we focus on environmental noise and ignore attacker-induced noise forms such as those caused by using evasive or polymorphic attack techniques.

The basic experiment involves running a single attack (ATK), iteratively for n times, against a target process running within an environment (ENV) and collecting symbolic data (DAT) in the form of system call traces using a monitor (MON).

- *Attack (ATK)* is a non-polymorphic, non-evasive attack. Every attack iteration performs the exact same sequence of operations represented by capital alphabets $A, B, C, D, \ldots$.
- *Monitor (MON)* observes the target process and produces sequential symbolic data in the form of an execution trace.
- *Symbolic data (DAT)* is sequence of records $a, b, c, d \ldots$ where one-or-more records map to some higher-level attack operation.
- *Environment (ENV)* is represented using a set of attribute value pairs further classified into the static and dynamic categories. Static includes attributes such as the operating system version, the distribution, the library version, the location of monitor and the version of monitor. Dynamic includes attributes such as the CPU usage profile, memory profile, disk usage profile and the network usage profile. All these attributes are specific to different environments.

The actual experiment is performed using two hosts: a virtualhost and a realhost. The *virtual host* container provides the environment (ENV). It runs the target process and the data monitor, along with other processes. The *realhost* runs the attack against the virtual host and controls the overall experiment. Having separate hosts ensures that the noise from the attacker's environment does not influence the target's environment.

3.2 Environmental Noise and Stability

The definition of *noise* is very subjective and we provide our definition and rationale in the next few paragraphs. There are atleast two aspects of noise that needs to be considered:

Noise depends on the data

Different sources of noise would have different effects over different types of data. For example, system-call trace data might not get affected by variations in CPU load but CPU usage data is influenced by variations in CPU load at different times of the day.

Noise depends on the detector

The simple reason here is that what seems like noise to one detector over some data might not be noise to the other over the same data. An analogy would be that Alice hearing Bob converse over the phone in his native-tongue would sound noisy to Alice but it is not noise for Bob.

For our purposes, the data is symbolic sequential data $a, b, c, d \ldots$ and detectors are assumed to use symbolic sequential data as their input for learning and detecting attacks.

Let a, b, c, d represent a stable attack signal. We define four types of unstable attack signals that can result due to environmental perturbations. The resulting instability can cause an attack signal to appear normal to the detector.

Unstable signal	Noise Type	Description
a, p, q, b, c, d	Addition	Extra symbols are added to original sequence
x, y, c, d	Replacement	Original symbols are replaced
a, c, d	Removal	Symbols are removed
a, c, b, d	Order Change	Order of symbols is changed
a_1, a_2, b, c, d	Split	Symbols are split into multiple symbols

3.3 Execution and Results

An initial experiment run was performed to capture the baseline set of symbols representing an attack. The basic experiment step described in Section 3.1 is then repeated under different conditions of environmental noise. All iterations of the attack are analyzed for the four types of unstable signals discussed in Section 3.2. In our experiments, we discovered the attack signal to manifest unstably due to addition, split, removal and replacement types of noise. We briefly discuss our preliminary results which demonstrate addition and split types of noise.

3.3.1 Unstable attack signal due to *addition*

The expected sequence of attack symbols were

```
...
time
time
...
```

One of the attack iterations produced the following sequence of calls.

```
...
time
stat64
write
time
time
....
```

The reason for addition here was that a file resource had changed causing the target process to execute a different set of system calls.

3.3.2 Unstable attack signal due to *split*

The original sequence contained the following system calls:

```
read  <====
stat64
stat64
stat64
stat64
clone
wait4
exit_group
```

Under conditions of load, it was observed that the single *read*, pointed to by the arrow, was split into two different reads as shown below.

```
read <====
stat64
stat64
stat64
stat64
clone
wait4
read  <====
exit_group
```

We thus see from above results that different instances of the same attack signal have manifested differently within the same environment.

4 Discussions

4.1 Reasons for attacks to manifest unstably

In the previous section, we showed that attacks do manifest unstably and here we present some of the possible reasons. For the case of host-based systems, an unstable attack manifestation may be:

an artifact of monitoring

Monitors capture data with differing levels of reliability under different conditions. For

example, the tcpdump monitor drops packets if the system is under heavy load. Monitors also have limitations which might make them capture data in ways not representative of the original system behavior. For example, the strace tool is known to have problems capturing system calls from processes forked with vfork(). Strace might also capture blocking system calls such as *send* as two calls spread over time.

an artifact of system execution

System behavior might change based on environmental conditions. For example, ordering of system call operations in multi-threaded applications might appear different for different executions, or a process might execute legitimate sequence of system calls which may have not been captured during training.

4.2 Applicability of results

Our experiments have demonstrated the effect of environmental noise using a simple host-based environment and for simple definitions of noise. We have shown that it is feasible for an attack to manifest unstably due to environmental noise. But, the results are not generic and still need to be verified across different types of environments. Our next steps in this regard are to investigate attack manifestation stability in large networks.

4.3 Evaluation of detector performance

Once we understand that an attack (A) manifests stably over some data (D) in environment (E), we can now proceed to make a confident assessment of a detectors performance. We can confidently assert that a good detector should always detect the attack A, over data D in environment E. Further, a failure to detect A would mean some problem with the detector's internal workings itself, according to our taxonomy in Figure 1.

4.4 Aspects of stability

We identify several different aspects of attack stability which need to be considered for a thorough understanding of the phenomena.

Stability with respect to data type

In this work, we considered sequential symbolic data which is categorical in nature. We need to investigate stability over different types of data such as continuous and time-series data.

Stability with respect to dimensionality

We considered single-dimensional symbolic data in this work. But, real world data is multidimensional with many different attributes. We hypothesize that it is possible for attacks to manifest stably in some dimensions while displaying instability across others. An understanding of this is essential for characterizing the performance of detectors and also for building better detectors.

Stability with respect to data granularity

Data can be monitored at different levels of granularity. For example, it is often the case that a monitor might capture only a subset of the system calls produced by a process. We expect an attack to show different stability with change in granularity.

5 Conclusions

Anomaly detection is an important technique in a defender's arsenal to protect against the latest breed of novel and sophisticated security attacks. It is well-understood that a sound evaluation methodology is required to make reliable assessment of a detector's performance and make it work across different environments.. In this work, we first laid out the evaluation space by describing various constructs that need to be considered for reliable evaluation of an anomaly detector. We then focused on understanding the impact of environmental noise on the stability of attacks. We started with the hypothesis that environmental noise can cause attacks to manifest unstably and demonstrated this with a simple experiment. Our immediate next steps involve investigating the phenomenon of attack manifestation carefully over different types of data, different types of environments and different detectors. Our larger goal is to build a comprehensive evaluation framework that can be used by researchers and evaluators to reliably evaluate anomaly detection technique in both the traditional cyber environments and cyber physical environments such as the Smart Grid.

Acknowledgements

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by Jet Propulsion Laboratory Year-Round Internship Program (JPLYIP) and the National Aeronautics and Space Administration. I sincerely thank my mentors Dr. Kymie Tan and Bryan Johnson for giving me the opportunity to pursue a summer internship at the Jet Propulsion Laboratory and supporting me through many odds. I especially thank Dr. Kymie Tan for making herself generously available despite her busy schedule, for always lending me a patient ear and for being enthusiastically supportive in all aspects of the research. I thank my advisor Dr. Clifford Neuman at the University of Southern California/Information Sciences Institute for being supportive of this internship opportunity. I cannot thank enough the remaining members of the Smart Grid cyber security team, Frank Kuykendall, Eric Rice, DJ Byrne, Brian Cox, Bradley Thomas, Thom McVittie and Mark McKelvin for the many stimulating discussions and for making the internship a truly memorable, enjoyable and fruitful experience. Last but not the least, I thank the Jet Propulsion Laboratory, California Institute of Technology for opening its doors to foreign nationals and allowing us the rare opportunity to experience cutting-edge research and engineering.

References

[1] CAPACCIO, T., AND BLISS, J. Chinese Military Suspected in Hacker Attacks on U.S. Satellites, 2011. http://www.businessweek.com/news/2011-10-27/

chinese-military-suspected-in-hacker-attacks-on-u-s-satellites.html.

[2] CHANDOLA, V., BANERJEE, A., AND KUMAR, V. Anomaly detection: A survey. *ACM Comput. Surv. 41*, 3 (July 2009), 15:1–15:58.

[3] CHANDOLA, V., MITHAL, V., AND KUMAR, V. Comparative evaluation of anomaly detection techniques for sequence data. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on* (dec. 2008), pp. 743 –748.

[4] FALLIERE, N., O MURCHU, L., AND CHIEN, E. W.32 Stuxnet Dossier v1.4, 2011. URL: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf.

[5] FORREST, S., HOFMEYR, S. A., SOMAYAJI, A., AND LONGSTAFF, T. A. A Sense of Self for Unix Processes. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 1996), SP '96, IEEE Computer Society, pp. 120–.

[6] GATES, C., AND TAYLOR, C. Challenging the Anomaly Detection Paradigm: a Provocative Discussion. In *Proceedings of the 2006 workshop on New security paradigms* (New York, NY, USA, 2007), NSPW '06, ACM, pp. 21–29.

[7] LEE, W., AND XIANG, D. Information-theoretic Measures for Anomaly Detection. In *Proceedings of the IEEE Symposium on Security and Privacy* (2001), pp. 130 –143.

[8] SOMMER, R., AND PAXSON, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *2010 IEEE Symposium on Security and Privacy (SP)* (May 2010), pp. 305 –316.

[9] WAGNER, D., AND SOTO, P. Mimicry attacks on host-based intrusion detection systems. In *Proceedings of the 9th ACM conference on Computer and communications security* (New York, NY, USA, 2002), CCS '02, ACM, pp. 255–264.