available at www.sciencedirect.com

ScienceDirect

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

Computers & Security

On the symbiosis of specification-based and anomaly-based detection

Natalia Stakhanova [a,*], Samik Basu [b], Johnny Wong [b]

[a] Faculty of Computer Science, University of New Brunswick Fredericton, NB E3B 5A3, Canada
[b] Department of Computer Science, Iowa State University Ames, IA 50011, USA

ARTICLE INFO

Article history:
Received 25 June 2009
Received in revised form
28 August 2009
Accepted 31 August 2009

Keywords:
Specification-based approach
Anomaly detection
Program behavior specification
Network monitoring
Intrusion detection

ABSTRACT

As the number of attacks on computer systems increases and become more sophisticated, there is an obvious need for intrusion detection systems to be able to effectively recognize the known attacks and adapt to novel threats. The specification-based intrusion detection has been long considered as a promising solution that integrates the characteristics of ideal intrusion detection system: the accuracy of detection and ability to recognize novel attacks. However, one of the main challenges of applying this technique in practice is its dependence on the user guidance in developing the specification of normal system behavior. In this work, we present an approach for automatic generation of specifications for any software systems executing on a single host based on the combination of two techniques: specification-based and anomaly-based approaches. The proposed technique allows automatic development of the normal and abnormal behavioral specifications in a form of variable-length patterns classified via anomaly-based approach. Specifically, we use machine-learning algorithm to classify fixed-length patterns generated via sliding window technique to infer the classification of variable-length patterns from the aggregation of the machine learning based classification results. We describe the design and implementation of our technique and show its practical applicability in the domain of security monitoring through simulation and experiments.

1. Introduction

The rapid increase in the number, sophistication and impact of computer attacks makes the computer systems unpredictable and unreliable, emphasizing the importance of intrusion detection ability to correctly recognize known attacks and identify new threats.

Typically, intrusion detection refers to a variety of techniques for detecting attacks in the form of malicious and unauthorized activities. There are three broad categories of detection approaches (Sekar et al., 2002) (a) misuse-based (b) anomaly-based and (c) specification-based. Misuse-based technique relies on pre-specified attack signatures, and any execution sequence matching with a signature is flagged as abnormal. An anomaly-based approach, on the other hand, depends on automatic classification of executions as normal patterns, and any deviation from normal patterns is classified as malicious or faulty. Unlike misuse-based detection, anomaly-based techniques can detect previously unknown abnormalities. However, anomaly-based approaches rely on statistical or machine learning classification techniques which can only classify (usually) pre-specified, fixed-length behavioral patterns, and suffer from the disadvantage of a high rate of false positives (Lazarevich et al., 2003). Specification-based techniques operate in a similar fashion to anomaly-based method detecting deviations from the

* Corresponding author.
E-mail address: nstakhanova@gmail.com (N. Stakhanova).
0167-4048/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2009.08.007

specified legitimate system behavior. However, as opposed to anomaly-based detection mechanism, specification-based approach requires user guidance in developing model of valid program behavior in a form of specifications. This process can handle variable-length sequences and is more accurate than anomaly-based techniques, but it can be prohibitively tedious and error-prone due to reliance on level of user-expertise.

1.1. *Problem statement*

The research in intrusion detection field has been mostly focused on anomaly-based and misuse-based detection techniques for a long time. This is due to the ability of anomaly detection approach to recognize novel attacks and predictability of misuse-based detection models (Innella, 2001; Kemmerer and Vigna, 2002). Although the specification-based detection has been less favored due to the inherent difficulty of developing specifications, it effectively combines the advantages of the other two approaches: the accuracy and the ability to detect new attacks, while avoiding their shortcomings. Sekar et al. (2002) emphasized another advantage of specification-based approach, its ability to accommodate the variable-length patterns that naturally represent the system behavior. (Sekar et al., 2001) showed that the traditional anomaly detection models employing machine learning based *n-gram* techniques may be error-prone in certain cases due to inability to classify patterns of variable-length.

Although the strengths of specification-based approach are obvious, the question of these benefits' applicability in practice still remains open. One of the main challenges in this context is the amount of effort required to develop normal behavioral specifications of large systems.

1.2. *Solution methodology*

In this paper, we explore the benefits of the specification-based approach in the security domain. Specifically, we analyze the capability of variable-length patterns to effectively represent system behavior and consequently, serve as a basis for system behavior specification.

In this work, we present an adaptive technique for automatic online learning and detection of abnormal program behavior through combination of two intrusion detection approaches: anomaly-based and specification-based detection. Such combination recognizes a known behavior of the system through the specifications of normal and abnormal patterns while classifying unknown behavior using a machine-learning algorithm.

Furthermore, in the absence of a pre-specified specification, the proposed approach can also be applied to automatically develop specifications using the classification results of machine-learning algorithms. Instead of manually developing all possible variable-length patterns of a legal system behavior, we use the machine-learning algorithm to classify fixed-length patterns as normal and/or abnormal, and then appropriately combine these classified patterns to create normal or abnormal behavioral specifications of variable-length. This technique of obtaining feedback from the machine-learning algorithm to create specification can also be used to augment and adapt pre-existing possibly incomplete specifications of normal behavior.

To efficiently maintain the results of classification and detect variable-length known/classified patterns, we propose a novel data structure *EXtended ACTion graph* (`Exact`). `Exact` was initially introduced in (Stakhanova et al., 2006) for automatic caching of monitored patterns in IDS. This work extends our previous research and provides a detailed study of `Exact` algorithms along with theoretical analysis between `Exact` patterns and fixed-length patterns. Furthermore, we also present a prototype showing the practical applicability of `Exact` in IDS.

In essence, the structure stores the specification of the normal and abnormal behavior of the system. `Exact` appropriately combines multiple sequences classified by a machine-learning technique into variable-length patterns and memorizes them for future reference. In our framework, we have two `Exact` structures: one for storing normal patterns and the other for abnormal patterns. Monitored sequences are classified using `Exact`, and the machine-learning algorithm is only invoked if necessary, i.e., if the monitored sequence does not match with any patterns stored in `Exact`. The following summarizes the contributions of this work:

1. *Adaptive detection of anomalies*. We employ a machine learning approach to classify unknown program behavior and memorize it for future reference, thus, effectively allowing the intrusion detection system to adapt to previously unknown normal and abnormal behaviors.
2. *Automatic development of specifications*. While the machine-learning technique automatically classifies fixed-length patterns, `Exact` caches the results of classification as variable-length sequences.
3. *Novel specification of behavior*. `Exact` allows compact and precise representation of variable-length sequences as specifications of normal and abnormal behavior.
4. *Efficiency of our approach*. We describe efficient algorithms for insertion of new patterns into `Exact` graph and identification of existing patterns using `Exact` graph. We present a prototype implementation of our technique and provide simulation results showing its effective applicability in the setting of system call based intrusion detection.

1.3. *Organization*

The remainder of the paper is organized as follows. A brief overview of related work is given in Section 2. Section 3 presents the `Exact` structure. Section 4 discusses the design and prototype implementation of our intrusion detection framework. Experimental results based on simulation and system prototype are given in Section 5. Section 6 concludes the paper with the discussion of future work.

2. Related work

The benefits of the specification-based approach have been noted by many researchers. The first specification-based models were introduced by Ko et al. (Ko et al., 1994, 1997, 2001;