Knowledge-Based Systems 106 (2016) 242–250

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Human error tolerant anomaly detection based on time-periodic packet sampling☆

Masato Uchida

Department of Information and Communication Systems Engineering, Faculty of Engineering, Chiba Institute of Technology, Narashino-shi, Chiba, 275-0016, Japan

ARTICLE INFO

Article history:
Received 19 November 2015
Revised 27 April 2016
Accepted 25 May 2016
Available online 26 May 2016

Keywords:
Anomaly detection
Human error
Packet sampling

ABSTRACT

This paper focuses on an anomaly detection method that uses a baseline model describing the normal behavior of network traffic as the basis for comparison with the audit network traffic. In the anomaly detection method, an alarm is raised if a pattern in the current network traffic deviates from the baseline model. The baseline model is often trained using normal traffic data extracted from traffic data for which all instances (i.e., packets) are manually labeled by human experts in advance as either normal or anomalous. However, since humans are fallible, some errors are inevitable in labeling traffic data. Therefore, in this paper, we propose an anomaly detection method that is tolerant to human errors in labeling traffic data. The fundamental idea behind the proposed method is to take advantage of the lossy nature of packet sampling for the purpose of correcting/preventing human errors in labeling traffic data. By using real traffic traces, we show that the proposed method can better detect anomalies regarding TCP SYN packets than the method that relies only on human labeling.

1. Introduction

Anomaly detection is the process of finding patterns in current network traffic that do not conform to legitimate (normal) behavior. The nonconforming patterns are called anomalies. Anomalies such as worms, port scans, denial of service attacks, spoofing, etc. seriously affect the operation and normal use of the network and may cause an enormous waste of network resources and economic loss. Consequently, anomaly detection has become an important issue in network monitoring and network security [2–7].

The design of an anomaly detection method usually relies on a baseline model describing the normal behavior of network traffic. An alarm is raised if a pattern in the current network traffic deviates from the baseline model. The baseline model is often trained using normal traffic data extracted from traffic data for which all instances (i.e., packets) are manually labeled by human experts in advance as either normal or anomalous. However, since humans are fallible, some errors are inevitable in labeling traffic data. Therefore, to achieve an efficient anomaly detection system, a method must be developed that extracts normal traffic data required for training the baseline model in a manner that is tolerant to human error in labeling traffic data.

In this paper, we have developed two methods that employ time-periodic packet sampling in conjunction with human labeling to solve this problem. That is, the two proposed methods employ time-periodic packet sampling to assist human experts in extracting normal traffic data required for training the baseline model. Since time-periodically sampled traffic contains a higher ratio of normal packets than the original traffic data [8], it is promising to employ time-periodic packet sampling to reduce the impact of human errors in labeling traffic data. Note that the two proposed methods are practically useful because they can reduce the effort human experts spend extracting normal traffic by using time-periodic packet sampling that has very low processing complexity. The difference in the two proposed methods is the operational order of human labeling and time-periodic packet sampling. The first method employs packet sampling after human labeling to correct human errors in labeling traffic data. That is, the first method makes primary use of human cognition to label traffic data and then secondary use of time-periodic packet sampling to correct human errors in the labeling process. This method is called the *ls*-method (labeling-and-sampling method) in this paper. The second method employs time-periodic packet sampling before human labeling to prevent human errors in labeling traffic data. That is, the second method makes primary use of time-periodic packet sampling to make cleaner traffic data that contains a higher ratio of normal packets than the original (unlabeled) traffic data and then secondary use of human cognition to label traffic data from the sampled traffic data. This method is called the *sl*-method

☆ A part of this paper appeared in the Proceedings of INCoS 2014 [1].
E-mail address: masato.uchida@ieee.org

http://dx.doi.org/10.1016/j.knosys.2016.05.050
0950-7051/© 2016 Elsevier B.V. All rights reserved.

(sampling-and-labeling method) in this paper. In both the *ls*- and *sl*-methods, the extracted normal traffic data is used for training a baseline model.

This paper is organized as follows. Section 2 briefly reviews related work on anomaly detection and packet sampling. Section 3 explains the fundamental idea behind the proposed method. Section 4 describes the experimental results obtained using actual traffic traces. Section 5 concludes the paper with a summary of the key points. The main differences between this paper and its original version [1] are the discussion about the ensemble-based anomaly detection given in Section 3.6 and the additional experimental results given in Section 4.4.

2. Related work

2.1. Intrusion detection

The process of securing a network infrastructure by scanning the network for suspicious activities is generically referred to as intrusion detection. The approaches to intrusion detection can be roughly classified into two categories: signature detection and anomaly detection.

2.1.1. Signature detection

In signature detection, the most widely deployed and commercially viable approach to detecting intrusions, the detection system identifies specific traffic patterns by matching the audited traffic data against the signatures of known attacks. The signatures are usually provided by human experts who investigate from the port number in the packet header to a specific byte sequence in the payloads of a series of packets. Snort [9] and Bro [10] are well-known open source systems that use signature detection. One benefit of this approach is that, once a signature database has been established, known attacks can be reliably detected with a low false positive rate. However, an alarm is not raised for attacks that are not registered in the database. For complete protection, the detection system must have a signature database containing all possible attacks, and the database must be manually updated whenever a new type of attack is discovered. Before such an update is made, the system is vulnerable to the new attack, meaning that the database must be frequently updated.

2.1.2. Anomaly detection

In anomaly detection, a baseline model is built for describing the normal behavior of network traffic. An alarm is raised if a pattern identified in the audited traffic data deviates from the baseline model. Unknown attacks can thus be detected because their behavior will deviate from the baseline model. Another benefit is that the anomaly detection approach is potentially easier to maintain than the signature detection approach because we do not need to update any signature records. Although false alarms are inevitable, the two benefits make the anomaly detection approach a promising area of research, and a number of methods based on this approach have been proposed [2–4].

We are interested in the anomaly detection approach, in which a baseline model is conventionally trained using normal traffic data extracted from labeled traffic data, where the label associated with an instance (i.e., a packet) denotes whether the instance is normal or anomalous. The labeled data is usually made by human experts. The basic problem with this is that errors in labeling traffic data are inevitable because humans are fallible. Therefore, we have developed two methods to reduce the impact of human errors in labeling traffic data. The fundamental idea behind the proposed methods is to take advantage of the lossy nature of packet sampling to correct/prevent human errors in labeling traffic data. Since time-periodically sampled traffic contains a higher ratio of normal packets than the original traffic data [8], it is promising to employ time-periodic packet sampling to reduce the impact of human errors in labeling traffic data. However, we previously [8] did not discuss how time-periodic packet sampling and erroneous human labeling should be used in combination in order to improve performance in anomaly detection. In this paper, we show that the two proposed methods (the *ls*- and *sl*-methods) can reduce the impact of human errors in labeling traffic data on the basis of theoretical analysis and experiments using actual traffic traces.

2.2. Packet sampling

Packet sampling has been attracting more and more attention as a way to minimize the resources needed to monitor traffic passing through high-speed backbone routers [11]. Modern routers already incorporate this technique, e.g., sFlow [12] and NetFlow [13]. Moreover, the Packet Sampling (PSAMP) Working Group [14] of the Internet Engineering Task Force (IETF) has standardized packet sampling techniques.

Although packet sampling provides greater scalability for network measurements [11,15–18], it makes inferring the original traffic characteristics much more difficult and biased because it is inherently lossy. For example, Kawahara et al. [19] showed that network anomalies generating a large number of small flows, such as network scans or SYN flooding, become difficult to detect during packet sampling. Therefore, they proposed a method that spatially partitions monitored traffic into groups for improving detection accuracy. On the other hand, Bartos et al. [20] proposed an intelligent flow sampling method to mitigate the negative impact of packet sampling for network security.

In previous work [8], we looked at these drawbacks of packet sampling from a different perspective. That is, we expected that the sampled (unlabeled) traffic would be favorably biased to the normal traffic by skipping the periods in which burst anomalies occur. This approach differs from that of other research on packet sampling in which the intent was to reduce the bias of the sampled traffic. We have confirmed that this conjecture holds true for time-periodically sampled traffic by analyzing actual traffic traces. In addition, we have found that a baseline model trained using time-periodically sampled (unlabeled) traffic data performs comparably for detecting anomalies regarding TCP SYN packets to one trained using manually labeled traffic data. That is, although the bias of sampled traffic is problematic for inferring the characteristics of the underlying traffic, this bias is not a drawback when the time-periodically sampled traffic, which contains a lower ratio of anomalous packets than the original traffic, is used to train a baseline model for anomaly detection.

3. Proposed methods

3.1. Procedure of proposed methods

We propose two methods (the *ls*- and *sl*-methods) that use time-periodic packet sampling to reduce the impact of human errors in labeling traffic data. The difference in the two proposed methods is the operational order of human labeling and time-periodic packet sampling (see Fig. 1). In the *ls*-method, human labeling is performed on the original traffic data first, where the label information may include some errors. Then, time-periodic packet sampling is performed on the labeled traffic data. The labeled-and-sampled packets are extracted eventually if they are labeled “normal”, otherwise the sampled packets are discarded. The *ls*-method is intended to correct human errors in labeling traffic data. In the *sl*-method, time-periodic packet sampling is performed on the original traffic data first. Then, human labeling