The Effects of Algorithmic Diversity on Anomaly Detector Performance

Kymie M.C. Tan and Roy A. Maxion
kmct@cs.cmu.edu and maxion@cs.cmu.edu

Dependable Systems Laboratory
Computer Science Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 / USA

Abstract

Common practice in anomaly-based intrusion detection assumes that one size fits all: a single anomaly detector should detect all anomalies. Compensation for any performance shortcoming is sometimes effected by resorting to correlation techniques, which could be seen as making use of detector diversity. Such diversity is intuitively based on the assumption that detector coverage is different – perhaps widely different – for different detectors, each covering some disparate portion of the anomaly space. Diversity, then, enhances detection coverage by combining the coverages of individual detectors across multiple sub-regions of the anomaly space, resulting in an overall detection coverage that is superior to the coverage of any one detector. No studies have been done, however, in which measured effects of diversity amongst anomaly detectors have been obtained.

This paper explores the effects of using diverse anomaly-detection algorithms in intrusion detection. Experimental results indicate that while performance/coverage improvements can in fact be effected by combining diverse detection algorithms, the gains are not the result of combining large, non-overlapping regions of the anomaly space. Rather, the gains are seen at the edges of the space, and are heavily dependent on the parameter values of the detectors, as well as on anomaly characteristics.

Based on this study, defenders can be provided with knowledge of how combinations of diverse, sequence-based detectors behave to effect detection performance superior to that of a single detector.

1 Introduction

Reliable, accurate and fast anomaly-based intrusion detectors capable of detecting known attacks, novel attacks and instances of insider threat over varying environments and data sets, remains as elusive today as it did two decades ago when Jim Anderson [1], and subsequently Dorothy Denning [4], proposed the idea. High false-alarm rates [3], inconsistency of detector performance [9, 20], and the inadvertent incorporation of intrusive behavior into a detector's concept of normal behavior (possibly causing the detector to miss the intrusion [11]), are only some examples of the numerous problems associated with the use of anomaly detectors in intrusion detection today.

Despite these many problems, anomaly detection remains, arguably, the most promising technique for detecting more insidious, and potentially more destructive, malicious incidents such as novel attacks and instances of insider threat. Such incidents are difficult to detect because they typically do not constitute a clear violation of security protocols and often lack clear, reliable signatures to facilitate their detection.

It is interesting to observe that despite the variety of anomaly detectors currently present in the intrusion detection literature, there appears to be an implicit assumption that a single anomaly detection algorithm is all that is required to detect intrusions or attacks on any given system. This assertion is supported by two further observations. First, intrusion detection systems claiming to perform anomaly detection typically employ only one kind of anomaly detection algorithm, e.g., [7, 8, 9, 18]. There is, however, no evidence to suggest that a single anomaly detector will be sufficient for a given intrusion detection task. No studies to date show that the kinds of anomalies that arise as manifestations of attacks are actually the kinds of anomalies that are detected by any given detector.

Second, of the studies that compare more than one anomaly detector (e.g., [10, 20]), the results of the respective efforts have been to determine the "best" single anomaly detection strategy for a data set. There are currently no studies, of which the authors are aware, acknowledging the possibility that effective intrusion detection may not necessarily be afforded by choosing the single, best-performing anomaly detector, but rather, by combining anomaly detectors such that the weaknesses of one may be compensated for by the strength of another, thereby taking advantage of diversity in detection algorithms.

Littlewood and Strigini [5] noted a renewed interest in using diversity for security. However, they also noted an absence of strategies by which to choose amongst diverse designs and by which to evaluate the effectiveness of the designs once selected. The present work takes inspiration from these observations, and examines how such choices can be made in anomaly detection. Namely, how can one make an informed choice amongst a set of anomaly detectors in a way that promotes improved detector performance? How can detectors be chosen such that their combined performance results in a net improvement? The evaluation strategy presented here enables one to study the effects of diversity on anomaly detection performance. The results of the evaluation describe the operational characteristics of a detector, providing a basis upon which to select amongst diverse detector designs. It also provides knowledge regarding the effects of combining more than one detector.

2 Background and related work

Anomaly detection can be regarded as simply a classification decision about an object or event: it is either anomalous or it's not. Although other researchers in the intrusion detection literature have extended the definition of anomaly detection to include causality (e.g., "anomaly detection attempts to quantify the usual or acceptable behavior and flags other irregular behavior as potentially intrusive"[11]), the present study views the attribution of cause as a separate decision and a separate problem to be solved. This latter position was adopted because it was deemed more important that the primary, and often only, capability of an anomaly detector, i.e., the detection of anomalies, be evaluated first. Subsequent evaluations can be performed to assess a detector's ability to identify those events that it does not directly detect, but that may have caused the detected anomalies (e.g., attacks and intrusions). An *anomaly-based intrusion detection system* therefore refers to a system that employs an anomaly detector as a component, and also attempts to link the detected anomalies to causal mechanisms such as attacks. It is interesting to note that at present, the link between anomalies and attacks is only an assumption [4].

It is possible for anomaly detectors to be blind to certain types of anomalous patterns. For example, an anomaly detector that does not employ probabilities such as Stide [20], cannot possibly detect an attack that manifests as a rarely occurring sequence. If detectors can be blind to certain types of anomalous patterns, it is not unreasonable to ask whether the detector can also be blind to those anomalous patterns that *are* the manifestations of attacks. [16] and [19] have shown that attacks may manifest, or even be manipulated to manifest, as normal behavior or as anomalous events that are invisible to a given anomaly-based intrusion detection system. The detection of attacks that manifest as normal behavior is obviously beyond the scope of an anomaly detector; however, the detection of attacks that do manifest as anomalous events is not out of scope. Figure 1 shows the necessary steps for determining whether or not an anomaly detector was successful at detecting an attack. The evaluation procedure described in this paper is focused on the last two issues, D and E, where the abilities specific to the anomaly detector itself are studied.

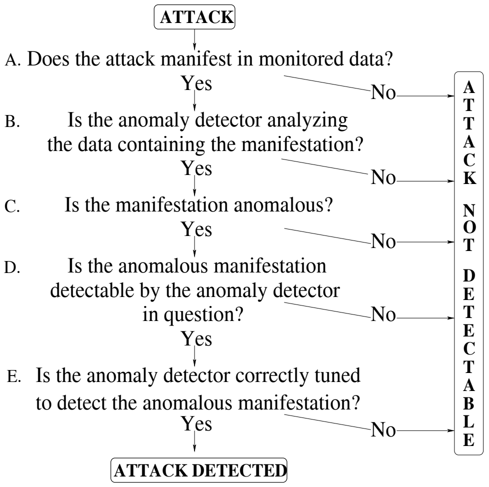
ATTACK
A. Does the attack manifest in monitored data?
Yes
No
B. Is the anomaly detector analyzing the data containing the manifestation?
Yes
No
C. Is the manifestation anomalous?
Yes
No
D. Is the anomalous manifestation detectable by the anomaly detector in question?
Yes
No
E. Is the anomaly detector correctly tuned to detect the anomalous manifestation?
Yes
No
ATTACK DETECTED
ATTACK NOT DETECTABLE

Figure 1: Determining the intrusion detection capability of an anomaly detector.

There are currently no studies centered on the *anomaly detection* capabilities of anomaly detectors, i.e., the kinds of anomalous events that a given anomaly detector actually detects, and how well. There are also currently no studies that employ such knowledge to examine the performance effects of diverse anomaly detection algorithms. The issue of diversity and security, however, was addressed in [5]. In that paper, the authors noted a growing awareness of diversity as a potentially valuable tool in the security community, but pointed out that none of the papers they examined addressed issues such as choosing amongst different diverse designs or evaluating the effectiveness of a selected design.

Previous work that attempts to evaluate anomaly-based intrusion detection systems does so without the diagnostic processes necessary to establish a factual link between attacks deployed and anomalies detected, i.e., issues A, B, and C of Figure 1. The evaluation procedure predominantly used in the literature for anomaly-based detection systems can be summarized in the following way [7, 9, 10, 12, 20]. Sets of *normal data* (obtained in the absence of intrusions or attacks) and *intrusive data* (obtained in the presence of intrusions or attacks) are collected. The anomaly-based intrusion detection system is trained on the normal data, and then tested on test data that contains either intrusive data only or some mixture of normal and intrusive data. The success of the detection algorithm is typically measured in terms of hit, miss, and false alarm rates, the ideal result being 100% hits, and 0% misses and false alarms.

This procedure is based on 3 unsupported assumptions:

1. An intrusion *will* manifest in data being analyzed by the anomaly-based intrusion detection system;
2. The manifestation is an anomalous event;
3. The anomaly detector within an anomaly-based intrusion detection system is capable of detecting the specific kinds of anomalous events that may be caused by the attacks of interest.

These assumptions make it impossible to determine, on the basis of the aforementioned evaluation procedure, whether or not a given anomaly detector is successful at detecting the attack. This is because there are three possible explanations for the results:

1. The anomaly detector is successful at detecting the anomalous manifestation(s) of the attack;
2. The detector detected anomalies that were the not the result of the attack;
3. The detector detected anomalies that resulted from an interaction between attack and non-attack events.

In short, all that can be determined from such results is that a set of anomalies was detected in the data, some or none of which may have been the anomalous manifestation(s) of the attacks of interest.

To assess the intrusion detection capabilities of an anomaly-based intrusion detection system, it is necessary to address each of the three assumptions listed above. To assess the capabilities of the anomaly detection component alone, however, it is only necessary to address the third assumption, as will be done by the present study. The evaluation procedure for anomaly-based intrusion detectors described above is set aside by this study because of the ambiguities inherent in the results and because it does not evaluate the capabilities of the anomaly detector with respect to the event it directly detects – the anomaly.

3 Hypothesis and claims

If it is true that any single anomaly detector can be applied to a detection task (because all anomaly detectors are equally capable of detecting anomalies, as has been presumed in the literature; see Section 1), then identical detection coverage is expected from each of the several anomaly detectors evaluated in this study, and the impact of employing diverse detection algorithms is minimal.

If differences are exhibited in detection coverage, then it can be concluded that not all anomaly detectors are equally capable; as a consequence, a single anomaly detector may not necessarily be sufficient for a given detection task, arguing for the use of diverse detection systems. The hypothesis is, therefore, that all anomaly detectors are equally capable of detecting anomalous events that may arise as manifestations of attacks or intrusions.

The claims made in this paper are also the lessons learned at the completion of the study, namely that (1) anomaly detectors designed to detect unequivocally anomalous events can be completely blind to these events; (2) diversity in detection methods has a significant effect on anomaly detection performance; (3) diversity in detection methods is manifested as differences in the conditions under which anomalous events can be detected; and (4) these conditions are affected by the characteristics of the anomalous event and by detector parameter values.

4 Approach

The approach taken by this study involves deploying a set of diverse, sequence-based anomaly detectors on carefully constructed synthetic data into which is injected a clearly defined, unequivocally anomalous event, detectable by all the chosen detectors. This approach was adopted for reasons that can be more clearly explained by addressing each of these three points: (1) the decision to use clearly defined anomalous events, (2) a set of sequence-based anomaly detectors, and (3) synthetic data.

4.1 Anomaly-based evaluation

An anomaly-based evaluation approach was adopted for two reasons. First, anomaly detectors do not detect attacks or faults unless they manifest as anomalies detectable by a given anomaly detector. It makes sense, therefore, to evaluate an anomaly detector's performance with respect to what it is designed to detect (anomalies) and not to what it may indirectly detect.

Second, the way in which an anomaly detector defines or perceives anomalies may not necessarily coincide with the ways in which anomalies naturally occur in data or with the kinds of anomalies that are manifestations of attacks. For example, an anomaly detector may be designed to detect foreign sequences, i.e., sequences that do not exist in the training data; however, in natural data, foreign sequences may exhibit unforeseen characteristics that interfere with the detector's abilities; this was in fact observed in the results and is discussed further in Section 7. Indeed, intrusion detection effectiveness for anomaly detectors can be described as a measure of the disparity between the kinds of anomalies that are the manifestations of attacks, and the kinds of anomalies that are detectable by the given anomaly detector. The greater the disparity, the worse the intrusion detection capability of an anomaly detector.

One may question whether the anomaly used in this study, the minimal foreign sequence described in Section

5.1, is of any significance in the real world, i.e., does this type of anomaly actually occur in natural data? Natural data was found to be replete with minimal foreign sequences of varying lengths. This is documented in [17] where datasets collected from various computer systems were analyzed; numerous instances of the minimal foreign sequence were found in the intrusive traces.

4.2 Sequence-based anomaly detectors

A set of sequence-based detectors comprises the element of diversity that is the basis of this study. The chosen detection algorithms can be described as "diverse" in that they were designed, created and deployed by different researchers for different projects, and diverse in the methods they use to effect anomaly detection.

The primary consideration guiding the choice of detectors was one of experimental control. Anomaly detectors may be diverse in a number of ways. Detectors may, for example, vary in the way they consume data – fixed-length sequences, single events, variable length sequences, etc.; detectors may also vary in the way by which they determine an event to be anomalous, by the way which normal behavior is modeled, and so forth. Choosing a set of anomaly detectors that are diverse in several dimensions would make it difficult to isolate the specific kind of diversity effecting detection performance or detection failure. It would also make it difficult to attribute observed detection performance to the effects of diversity as opposed to other effects. For this reason, the substance of diversity in this study is constrained to only one, and arguably the most important, aspect of the chosen anomaly detectors – the methods by which deviations from normal behavior are measured.

The anomaly detectors selected for this study can be described as consisting of three general components:

1. A mechanism for modeling normal behavior;
2. A metric or method for measuring deviations from the model of normal behavior;
3. A thresholding mechanism for determining whether the detected deviation is significant enough to label the event as anomalous with respect to normal behavior.

Whereas the detectors are diverse in the second of these components, they are invariant in the other two. The basic event being analyzed is the same for all four selected detectors – the fixed-length sequence. All four detectors are expected to be able to detect the same anomaly, a foreign sequence. The thresholding mechanisms for all four detectors are set by the user, and as such are controlled by ensuring that the definitions of hits and misses are consistent across all four detectors. This last issue will be addressed further in Section 5.5.

4.3 Synthetic data

Natural data was not used in this study because it was necessary to ensure that the data upon which the detectors were to be deployed did not contain confounding elements that can undermine the fidelity of the final results. To this end, this study employed synthetic data because it provided the control necessary for constructing the defined anomalous event, for constructing training and background test data that were free of spurious, naturally occurring anomalies, and for enabling an injection procedure that kept the character of the anomalous event and the background data intact.

5 Experimental methodology

The effect of diversity on detector performance is examined by focusing narrowly on the abilities of a set of chosen anomaly detectors to detect a single, clearly defined, and unequivocally anomalous event. The following list provides an overview of the experimental methodology designed to support this intent; subsequent sections describe the experiment in detail, giving the rationales and motivations behind the decisions made for each stage of the experiment.

1. Define the anomaly.
2. Select the detectors.
3. Synthesize the training (normal) data.
4. Synthesize the test data.
   (a) Synthesize the background data.
   (b) Synthesize the anomaly, and inject the anomaly into the background data to create the final corpus of test data.
5. Deploy the anomaly detectors on the synthesized training data and on the test data.
6. Analyze the results.

5.1 The anomaly

The anomalous event used in this study is referred to as a minimal foreign sequence (MFS). A foreign sequence can be described as a sequence of length N where each individual element within the sequence is a member of the training-set alphabet, but where the entire length-N sequence itself does not occur in the training data. A *minimal foreign sequence* is a foreign sequence with the property that all of its proper sub-sequences *do* exist in the normal data [15, 17]. Put simply, a minimal foreign sequence is a foreign sequence that contains within it no smaller foreign sequences.

The decision to employ only one anomaly type in this experiment was prompted by two reasons. First, for anomaly detectors that employ fixed-length sequences, the foreign sequence is exemplary of the kind of anomaly that should be detectable by all sequence-based anomaly detectors – unlike, for example, rare sequences. Rare sequences are detectable by some detectors, e.g., Markov-based detectors, but are not detectable by others, e.g., Stide and the Lane and Brodley (L&B) detector. Furthermore, the intrusion detection literature remains ambiguous about the "alarm-worthiness" or "anomalous-ness" of rare sequences [20].

Secondly, the use of a single anomaly type would more clearly illustrate the point that if there are wide variations in detection capabilities over a single, detectable anomaly, then the intrusion detection practice of deploying only one anomaly detection strategy in a given intrusion detection system may be prone to failure, arguing for the use of diverse detection strategies.

5.2 The sequence-based anomaly detectors

Four sequence-based anomaly detectors were examined in this study. Their selection was governed by the need to constrain diversity to only one aspect of the detector for experimental control. All four detectors analyze fixed-length sequences of categorical data, and conform to the generic description of an anomaly detector described in Section 4.2. Their diversity, however, lies in the manner in which they each determine the abnormality of a given sequence, i.e., their similarity metric. Where one algorithm may employ probabilistic concepts to determine such abnormality, another would establish abnormality merely as a difference in the ordering of elements within a sequence. The detectors examined in this study are a Markov model-based detector, a Neural Network-based detector, Stide, and the Lane and Brodley detector.

This section describes the similarity metrics for each of the four anomaly detectors under scrutiny. Normal behavior for all four detectors is acquired by sliding a detector window of fixed-length size (DW) across the training data, and storing the DW-sized sequences in a database.

Markov-based detector. The Markov-based anomaly detector [12, 18] employs the sequential ordering of events and conditional probabilities in its detection approach. For every fixed-length sequence of size DW obtained from the test data, the detector calculates the probability that the "$DW + 1$"st element will follow. The detector produces a score between 0 and 1 for each element in the test data stream beginning at the "$DW + 1$"st element. This score indicates the probability that the "$DW + 1$"st element followed the previous size-DW sequence, where 1 indicates highly improbable and 0 indicates normal (very probable).

Neural-network-based detector. The Neural-network-based anomaly detector [6] employs sequential ordering of events in its detection approach. The similarity metric for this detector is essentially embedded in the multilayer, feed-forward learning mechanism. Although it does not use explicit probabilistic concepts, the detector's learning algorithm is an approximation function that can be described as "mimicking" the effects of employing probabilistic concepts such as the conditional probabilities used by the Markov-based detector.

Stide detector. Stide [7, 20] is an anomaly detector that is completely dependent upon the sequential ordering of categorical elements in the data stream. The detector establishes whether every fixed-length sequence of size DW from the test data exists in the normal database of same-sized sequences. The value 0 is assigned to indicate that a matching normal sequence was found, and the value 1 is assigned to indicate otherwise. No direct probabilistic concepts, such as the calculation of frequencies or conditional probabilities, are employed by this algorithm.

Lane & Brodley detector. The Lane & Brodley detector (L&B) [13] is also completely dependent on the sequential ordering of elements in the data stream. For two fixed-length sequences of the same size, each element in one sequence is compared to its counterpart at the same position in the other sequence. Elements that do not match are given the value 0, and matching elements are given a score that incorporates a weight value. This weight value increases as more adjacent elements are found to match. The similarity metric produces a value between 0 and $DW(DW + 1)/2$, where 0 denotes the greatest degree of dissimilarity (anomaly) between the two sequences and $DW(DW + 1)/2$ denotes the greatest degree of similarity (identical sequences). No probabilistic concepts such as the calculation of frequencies or conditional probabilities are used by this detector.

5.3 The training data

The generation process of the evaluation dataset is documented in detail in [14] and [17]; hence, this section will only describe the characteristics of the evaluation data that are pertinent to the present experiment.

A training-data stream of 1,000,000 elements was constructed using a Markov-model transition matrix. Three parameters were chosen arbitrarily in this experiment: the sample size of 1,000,000 elements; the length of the minimal foreign sequences (denoted AS for *anomaly size*), which ranged from 2 to 9; and the definition of a rare sequence – a *rare* sequence is one with a relative frequency of less than 0.5% in the training data [20].

The alphabet size for the training data was 8. Although alphabet sizes in real-world data are higher than this and may influence, for example, the size of the set of possible sequences that populate the normal database, the alphabet size of the training data does not affect the synthesis of foreign sequences, nor does it affect a sequence-based detector's ability to detect foreign sequences.

Ninety-eight percent of the one-million-element data stream consisted of a repetition of the sequence "12345678." This characteristic of the data provided a consistent set of "noiseless" common sequences that were independent of sequence length, and that could be used to populate the background data, i.e., the test data without the injected anomaly. The remaining two percent of the training data contained rare sequences that were the result of a small amount of nondeterminism in the probabilities of the data generation matrix. These rare sequences were necessary for synthesizing the minimal foreign sequence anomaly composed of rare sub-sequences.

5.4 The test data

The test data were constructed in two stages. First the background data was created; after this, the anomalies were synthesized and injected into the background data to form the final corpus of test data.

5.4.1 The background data

The background data were composed of commonly occurring sequences found in the training data, i.e., a repetition of the sequence "12345678." This ensured that the background data was "clean." and contain no spurious, naturally occurring foreign or rare sequences that might confound the results of the experiment. Hence, any detector-window size ranging from 1 to the data sample size of a million, sliding over the background data, would only experience common sequences already present in the training data.

5.4.2 Creating and injecting the anomalies

The training data were used to create the anomalies – minimal foreign sequences of sizes 2 to 9, composed of rare sub-sequences. A rare sequence is simply defined to be a sequence with a relative frequency of 0.5% in the training data, a definition taken from previous work [20].

The decision to use rare sub-sequences was prompted by the expectation that both the Neural Network detector and the Markov detector should be able to respond to rare sequences. Although some detectors (such as Stide) do not have the ability to respond to rare sequences, they are nevertheless applied to anomalies having these characteristics, primarily to facilitate performance comparisons; i.e., all the detectors in question are evaluated on their ability to detect the same anomaly. Furthermore, it would be a point of interest to observe, and possibly quantify, how much more accuracy the ability to detect rare sequences actually confers upon the detection of foreign sequences so composed.

Sequences composed by concatenating short, rare sequences from the training trace are likely to be foreign, simply due to the improbability that a substantial number of rare sequences would appear in the training trace in the chosen order. It is easy to generate such sequences, and to verify their "foreign-ness" and minimality characteristics. These same characteristics, however, complicate the process of injecting the anomaly, which unfortunately remains somewhat of a brute force effort.

The sub-sequences within the composed anomaly tended to interact with the background data to produce spurious rare and/or foreign sequences. These unintended anomalies were most likely to occur at the boundaries where some elements of the injected anomaly and some elements of the background data combined within the sliding detector window to produce unintended foreign or rare sequences. Figure 2 provides examples of these "boundary sequences." Randomly injecting an anomaly into the background data is undesirable because of the high probability that a mixture of foreign or rare boundary sequences is introduced by such an injection strategy.

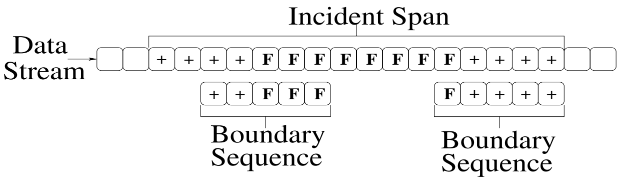

Figure 2: Boundary sequences with detector window of size 5 and foreign sequence of size 8. **F:** injected foreign sequence; $+$**:** elements of background involved in boundary sequences. The incident span comprises all 5-element sequences that contain at least one element of the anomaly.

Given a detector window of size DW and a minimal foreign sequence anomaly, a desirable injection procedure is one ensuring that all of the $2(\mathrm{DW}-1)$ sequences of length DW that can be composed at the boundary of the injection, i.e., sequences that contain some elements of the anomaly and some elements of the background data, are common sequences that exist in the training data. It must be ensured that no background data sequences or boundary sequences register as foreign or rare. If this is not possible for some location in the trace, a new anomaly must be produced as a replacement, and the process repeated.

The final suite of evaluation data contains one stream of training data and 8 streams of test data, where each test-data stream contains a single minimal foreign sequence whose length is selected from the range 2 to 9. This set of 9 data streams is then replicated for each detector-window length of 2 to 15. In total there are 112 test-data streams.

5.5 Detector deployment and scoring

Each detector was deployed on the suite of data created in the preceding section. For each minimal foreign sequence being detected, the length of the detector window was varied from 2 to 15. Processes occurring *after* the application of the similarity measure were ignored, e.g., Stide's locality frame count (LFC). Only a detector's intrinsic ability to detect a specific anomalous phenomenon was considered – not noise-suppression techniques like the LFC.

The results of the experiments are expressed in terms of hits, misses, and regions of detection blindness and weakness. When a detector window slides over an anomaly and encounters a boundary sequence, the interaction between the elements of the anomalous sequence and the background data will prompt the detector to produce a response that is influenced by the elements of the injected anomaly.

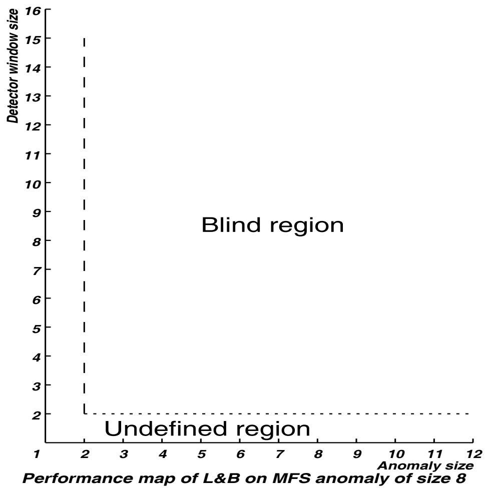

Figure 3: Detection coverage, L&B detector.

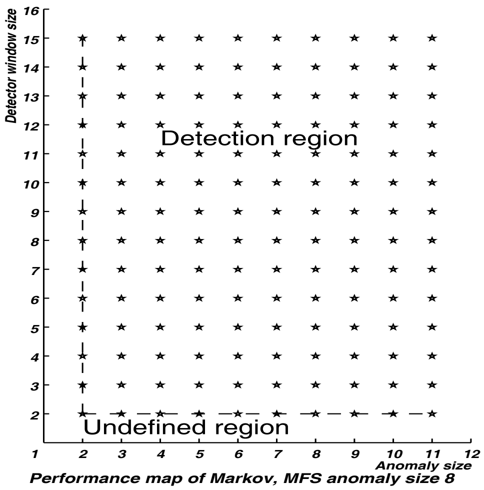

Figure 4: Detection coverage, Markov-based detector.

This issue is considered in the scoring process and resulted in the concept of the incident span [17], see Figure 2.

Allowing a detector's response to range from 0 (indicating completely normal) to 1 (indicating maximal abnormality), a detector is described as: *blind*, in the case where the detector response is 0 for *every* sequence of the incident span; *weak*, in the case where the maximum detector response registered in the incident span is greater than 0 and less than 1, indicating that something definitely abnormal has been seen; and *capable*, in the case where at least one detector response of 1 was registered in the incident span.

Binary detectors, such as the sequence-matching portion of Stide, are only capable of generating responses of 0 or 1; however, the Neural Network and the Markov-based detector can generate weak responses. Weak responses can be converted to binary responses by applying a threshold that converts responses below the threshold to 0 and others to 1. To facilitate fair comparisons among these detectors, the detection threshold was set to 1 for all detectors, recognizing only maximally anomalous (foreign) sequences as "hits."[1]

6 Results

The results from the experiment described above are displayed in four graphs. Figure 3 presents the detection capability for the L&B detector, Figure 4 for the Markov-based detector, Figure 5 for Stide, and Figure 6 for the Neural-Network-based detector.

The x-axis in each figure marks the increasing length of the minimal foreign sequence (MFS) injected into the test-data stream, and the y-axis charts the length of the detector window. Each star marks the length of the detector window (on the y-axis) required to detect a foreign sequence whose corresponding length is marked on the x-axis; the term "detect" specifically means that a maximum anomalous response occurred in the incident span. The areas that are absent of a star (blind regions) indicate that the foreign sequence whose corresponding length is marked on the x-axis was perceived by the detector as being a completely normal sequence.

Since the Markov-based detector utilizes the Markov assumption, i.e., that the next state is dependent only upon the current state, the smallest window length possible is 2. This means that the next expected, single, categorical element is dependent only on the current, single, categorical element. As a result, the y-axis marking the detector-window lengths in Figure 4 begins at 2. The same argument applies with the Neural-Network-based detector in the sense that this detector predicts the next categorical element based on the current categorical element; this makes 2 the smallest workable detector window length. Although it is technically possible to run Stide and the L&B detector using a detector window of length 1, doing so would produce results that do not include the sequential ordering of events, a property that comes into play with all the detector-window lengths that are larger than 1. This, together with the fact that there is no equivalent for either the Neural-Network-based detector and the Markov-based detector, argued against running Stide and the L&B detector with a window of 1.

The x-axis also begins at 2 in each figure. This is because the type of anomalous event upon which the detectors are being evaluated requires that a foreign sequence be composed of rare sequences. A foreign sequence of size 1,

[1] Detection thresholds are often used to determine "alarm-worthy" events. The maximum anomalous response will always register as an alarm regardless of where the detection threshold is set. An anomalous phenomenon generating such a response will never "disappear" or become a miss when the detection threshold is raised or lowered.

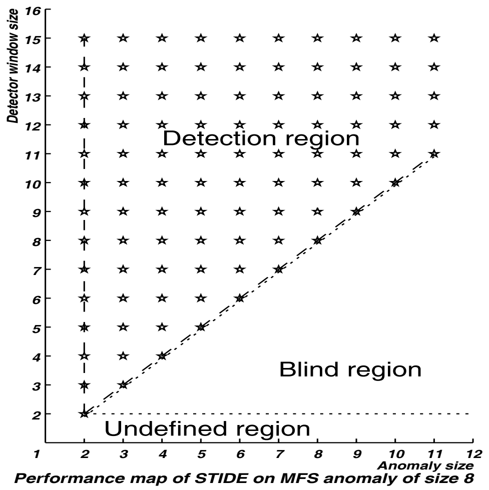

Figure 5: Detection coverage, Stide.

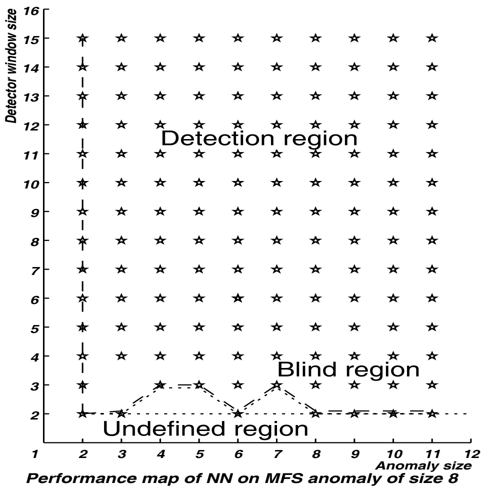

Figure 6: Detection coverage, Neural-Net-based detector.

therefore, will contain a single element that must be both foreign and rare at the same time, and this is not possible. As a consequence, all the result graphs show an undefined region corresponding to the anomaly size of 1.

7 Discussion

The results show that despite the fact that all the detectors analyze data in terms of sliding windows of fixed-length sequences, and that are all expected to detect foreign sequences, their differing similarity measures significantly affected their detection capabilities. There are four main points to note; these are operational details that must be considered when deploying these detectors.

First, there are regions of detection blindness. Even for an event as unequivocally anomalous as a foreign sequence, some sequence-based detectors are unable to detect its presence. The L&B detector, for example, will classify a minimal foreign sequence as a sequence close to normal, while Stide will classify that same sequence as an anomaly, but only when $DetectorWindow > AnomalySize$.

Second, the results show that the different similarity measures used by each detector significantly affect detection performance. In Stide's case, even though it is certain that there is a foreign sequence present in the data stream, this foreign sequence is only visible if the length of the detector window is at least as large as the length of the foreign sequence. The Markov-based detector, on the other hand, appears to have no such weakness. The foreign sequence in the data stream is visible to the Markov detector, even when the length of the detector window is smaller than the length of the foreign sequence. This can be attributed to the use of rare sequences in composing the foreign sequence. Under such a circumstance the use of conditional probabilities appears to have afforded the Markov detector an advantage.

Third, the results provide some knowledge of how to combine detectors to effect detection performance. Consider, for example, the observation that Stide will only detect foreign sequences, while the Markov-based detector will detect foreign sequences as well as a variety of rare sequences. It follows that if the Markov-based detector is deployed for intrusion detection purposes, it can only be expected to produce greater numbers of false alarms than Stide. This is because it will detect anything that manifests in the data stream as a foreign symbol (this can be seen as a foreign sequence of size 1), and various compositions of foreign and rare sequences. Stide on the other hand will only detect foreign sequences. A circumstance under which this knowledge may be useful would be, for example, when it is known that an attack typically manifests as a minimal foreign sequence, but the size of this foreign sequence is unknown (making Stide unreliable as the main detector since Stide would only detect such a manifestation if its detector window is set to at least the *known* size of the minimal foreign sequence). The Markov-based detector can be used to detect the manifestation of the attack itself while Stide can be used as a suppressive mechanism to reduce false alarms. Any alarms raised by the Markov-based detector, and not raised by Stide, may be ignored as false alarms; alarms raised by both Stide and the Markov-based detector are possible hits. Any alarm raised by Stide will also be raised by the Markov detector, because it is now known from the results of the evaluation that Stide's detection coverage is a subset of the Markov-based detector's coverage.

Although the combination of Stide and the Markov-based detector may produce performance improvements in the form of reduced false alarms, the combination of Stide

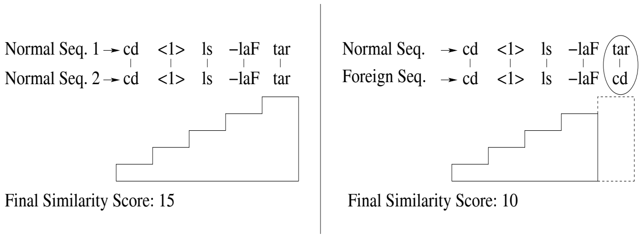
Normal Seq. 1 → cd <1> ls −laF tar
Normal Seq. 2 → cd <1> ls −laF tar
Final Similarity Score: 15
Normal Seq. → cd <1> ls −laF tar
Foreign Seq. → cd <1> ls −laF cd
Final Similarity Score: 10

Figure 7: Similarity calculation between two size-5 sequences. The step curve represents the weight contributed by each match. Note the detector responds with a value close to normal when presented with a foreign sequence.

and the L&B detector is not so agreeable. While both of these detectors use very different mechanisms by which to detect an anomaly, they both show a blindness in the same region of the performance chart. In this case, employing the two detectors to take advantage of diverse detection algorithms affords no improvement in performance.

The Neural-network-based detector, although using a different mechanism by which to effect detection, appears to be as good as the Markov-based detector. The only caveat with this detector lies in the reliability of detection. It is common knowledge that the performance of a multi-layer, feed-forward network relies on a balance of parameter values, e.g., the learning constant, the number of hidden nodes, and the momentum constant [21]. Some combinations of these values may result in weakened anomaly signals. In these cases, the setting of another parameter – the detection threshold – becomes critical for detecting particular events.

The fourth and final point is the observation that by charting the performance of the detectors with respect to the detection of minimal foreign sequences, it was possible to observe the nature of the gain achieved in detection performance between an algorithm that employs conditional probabilities and one that employs, for example, only sequence-matching schemes. This gain in detection ability is significant and is illustrated by the blind regions marked out in Figures 5 (Stide) and 3 (L&B).

A detail of concern is the inability of the L&B detector to detect a minimal foreign sequence anomaly even when the entire foreign sequence can be "seen" by the detector, i.e., $DW = AS$. The blindness in the L&B detector can be attributed to the similarity metric that the detector employs. The similarity metric for the L&B detector is biased in favor of matching adjacent elements. It is possible for a foreign sequence to match every element of a normal sequence except one. If the single, mismatching element is located at the very first or very last position of the foreign sequence, the detector returns a similarity value indicating that the foreign sequence is close to normal.

To illustrate the point more clearly, the left-hand diagram in Figure 7 depicts the similarity calculation between two identical sequences of size 5. The similarity value for these two identical sequences is $Sim_{max} = \sum_{i=1}^{DW} i = \frac{DW(DW+1)}{2} = 15$. This is the highest, most "normal" value that the similarity metric can produce for a detector window of size 5. However, when a normal and foreign sequence are compared, as shown in the right-hand diagram of Figure 7, the only difference between the two sequences is the last element, and the similarity value computes to $Sim_{weak} = \sum_{i=1}^{DW} i = \frac{DW(DW-1)}{2} = 10$. The slight dip in the similarity value from 15 to 10 is all that indicates the presence of the foreign sequence (the most anomalous result for this detector is the value 0). For the L&B detector to detect the foreign-sequence anomaly, the detection threshold must be set to the next most normal value, which is 10. If this is done, then all sequences from the test data that differ from a normal sequence by at least one element, will register as anomalous. This will raise the false alarm rate, which will get increasingly worse as the sequence length grows.

8 Conclusion

The hypothesis that all anomaly detectors are equally capable of detecting anomalies was tested in this study, and determined experimentally to be without support.

The sequence-based anomaly detectors examined here exhibit diversity primarily on the basis of their similarity metrics: the conditional probabilities of the Markov-based detector, the exact matches of Stide, the weighted matches of Lane & Brodley, and the function approximation strategy of neural nets. This diversity in similarity metrics leads to diversity in coverage of the detection space.

The Markov detector covers the entire space under consideration, but in doing so it is prone to false alarms. The Lane & Brodley detector, despite its previous application to masquerade detection, is blind across the entire space considered. Stide covers a subset of the Markov space, and is sensitive to conditions under which its detector window is shorter than the anomalous minimal foreign sequence being detected. The neural-net detector mimics the Markov detector; however, it is also highly dependent on the art of setting its tuning parameters.

Obtaining increased detection coverage is an obvious goal of exploiting diversity, but combining these detectors requires care, as the experimental results have shown. Stide, for example, will only detect foreign sequences, whereas the Markov-based detector will detect foreign sequences as well as a variety of rare sequences. Stide can be unreliable when an attack manifests as a foreign sequence, but the sequence size is unknown; in such cases the Markov-based detector can be used to detect the manifestation of the attack itself while Stide can be used as a suppressive mechanism to reduce false alarms.

In another example, combining Stide and L&B provides no detection advantage at all. Although each of these detectors uses a very different similarity metric, they each show blindness in the same region of the performance chart. In this case, diversity affords no improvement in detection performance (hits) for two reasons: (1) both detectors will be blind to the presence of an MFS anomaly when deployed with a detector window size that is less than the size of the anomaly; (2) when the detector window is set to be equal to the size of the anomaly, only Stide is capable of detecting an MFS anomaly when only the first or last element mismatched any normal sequence. No detection advantage is gained by the presence of the L&B detector in this case.

The experiments showed that certain anomalous sequences (MFSs) are not detectable under certain conditions by certain detectors. Since real-world datasets contain numerous instances of these anomalous sequences, the ramification for intrusion detection is that some anomaly detectors will be unable to detect attacks that manifest as minimum foreign sequence anomalies. Moreover, detectors that are able to detect attacks that manifest as MFSs can be rendered blind to these attacks by an incorrect choice of detector parameter values (e.g., detector window size).

The experiments reported here were based on the detectors' native abilities for detecting anomalies. The results maintain their validity when extended beyond synthetic data, into the real world; there is no difference between a minimal foreign sequence embedded in synthetic vs. natural data. But, future work must go farther; it must focus on the relationship between detectable anomalies and intrusive behaviors. It is critical to establish that the anomalies detected were caused by attacks, and not by more innocuous events. This is not a trivial task.

References

[1] James P. Anderson, "Computer Security Threat Monitoring and Surveillance", Technical report, James P. Anderson Co., Fort Washington, Pennsylvania, April 1980.

[2] Matt Bishop, Steven Cheung, and Chris Wee, "The Threat from the Net", 34(8) *IEEE Spectrum,* pages 56-63, August 1997.

[3] Herve Debar, Marc Dacier and Andreas Wespi, "Towards a Taxonomy of Intrusion-Detection Systems", *Computer Networks,* 31(8), pages 805-822, April 1999.

[4] Dorothy Denning, "An Intrusion-Detection Model", *IEEE Transactions on Software Engineering,* SE-13(2), pages 222-232, Feb. 1987.

[5] Bev Littlewood and Lorenzo Strigini, "Redundancy and Diversity in Security", In *Proceedings ESORICS 2004, 9th European Symposium on Research in Computer Security,*, pages 423-438, Sophia Antipolis, France, September 2004, Lecture Notes In Computer Science # 3193, Springer-Verlag, Berlin, 2004.

[6] Herve Debar, Monique Becker, and Didier Siboni, "A Neural Network Component for an Intrusion Detection System", In *Proceedings of the 1992 IEEE Computer Society Symposium on Research in Security and Privacy,* pages 240-250, 04-06 May 1992, Oakland, CA. IEEE Computer Society Press, Los Alamitos, CA.

[7] Stephanie Forrest, Steven A. Hofmeyr, Anil Somayaji, and Thomas A. Longstaff, "A Sense of Self for Unix Processes", In *Proc. 1996 IEEE Symp. on Security and Privacy,* pp. 120-128, 06-08 May 1996, Oakland, CA. IEEE Computer Security Press, Los Alamitos, CA.

[8] Kevin L. Fox, Ronda R. Henning, Jonathan H. Reed, and Richard Simonian, "A Neural Network Approach Towards Intrusion Detection", In *Proceedings of the 13th National Computer Security Conference,* pages 125-134, 01-04 October 1990, Washington DC.

[9] Anup Gosh, Aaron Schwartzbard and Michael Schatz "Learning Program Behavior Profiles for Intrusion Detection", In *Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring,* pages 51-62, 09-12 April 1999, Santa Clara, CA. USENIX Association, Berkeley, CA.

[10] Anup Gosh, James Wanken, and Frank Charron "Detecting Anomalous and Unknown Intrusions Against Programs", In *Proceedings of the 14th Annual Computer Security Applications Conference,* pages 259-267, Scottsdale, AZ, 07-11 December 1998. IEEE Computer Security Press, Los Alamitos, CA.

[11] Sandeep Kumar, "Classification and Detection of Computer Intrusions", Ph.D. Dissertation, Purdue University, West Lafayette, Indiana, August 1995.

[12] Somesh Jha, Kymie M. C. Tan, and Roy A. Maxion, "Markov Chains, Classifiers, and Intrusion Detection", In *Proceedings of the 14th IEEE Computer Security Foundations Workshop*, pages 206-219, 11-13 June 2001, Cape Breton, Nova Scotia, Canada.

[13] Terran Lane and Carla E. Brodley, "Sequence Matching and Learning in Anomaly Detection for Computer Security", In *Proc. of AAAI-97 Workshop: AI Approaches to Fraud Detection and Risk Management*, AAAI Press, pp. 43-49, 27-31 July 1997, Providence, RI.

[14] Roy A. Maxion and Kymie M. C. Tan, "Benchmarking Anomaly-Based Detection Systems", In *Proceedings of the International Conference on Dependable Systems and Networks*, pages 623-630, 25-28 June 2000, New York, NY. IEEE Computer Society Press.

[15] Roy A. Maxion and Kymie M. C. Tan, "Anomaly Detection in Embedded Systems", *IEEE Transactions on Computers*, 51(2), pages 108-120, February 2002.

[16] Kymie M. C. Tan; Kevin S. Killourhy, and Roy A. Maxion, "Undermining an Anomaly-Based Intrusion Detection System Using Common Exploits", In *Proceedings of the Fifth International Symposium on Recent Advances in Intrusion Detection (RAID-2002)*, Andreas Wespi, Giovanni Vigna and Luca Deri (Eds.), 16-18 October 2002, Zurich, Switzerland, pp. 54-73. Lecture Notes in Computer Science #2516, Springer-Verlag, Berlin, 2002.

[17] Kymie M. C. Tan and Roy A. Maxion, " 'Why 6?' Defining the Operational Limits of stide, an Anomaly-Based Intrusion Detector", In *Proceedings 2002 IEEE Symposium on Security and Privacy*, pages 188-201, 12-15 May 2002, Oakland, CA. IEEE Computer Society Press, Los Alamitos, CA.

[18] Henry S. Teng, Kaihu Chen, and Stephen C-Y Lu, "Security Audit Trail Analysis Using Inductively Generated Predictive Rules", In *Proc. of the Sixth Conference on Artificial Intelligence Applications*, pages 24-29, IEEE Service Center, Piscataway, NJ, March 1990.

[19] David Wagner and Paolo Soto. "Mimicry attacks on host-based intrusion detection systems", In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pp. 255-264, 18-22 Nov. 2002, Washington, DC. ACM Press, New York, NY, USA.

[20] Christina Warrender, Stephanie Forrest and Barak Pearlmutter, "Detecting Intrusions Using System Calls: Alternative Data Models", In *Proceedings 1999 IEEE Symposium on Security and Privacy,* pages 133-145, 09-12 May 1999, Oakland, CA. IEEE Computer Security Press, Los Alamitos, CA.

[21] J. Zurada. "Introduction to artificial neural systems," West Publishing Co., St. Paul, MN, 1992.