"Why 6?" Defining the Operational Limits of stide, an Anomaly-Based Intrusion Detector

Kymie M.C. Tan and Roy A. Maxion

November 2001
CMU-CS-01-158

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This research was supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under contracts F30602-99-2-0537 and F30602-00-2-0528.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA or the U.S. government.

Keywords: Anomaly, anomaly detection, detection coverage, evaluating anomaly detectors, stide.

Abstract

The detection of masqueraders and novel attacks are two of the more difficult problems facing intrusion detection systems. While anomaly-based intrusion detection approaches appear to be among the most promising techniques for dealing with these problems, confidence in the detection results requires precise knowledge of the detector's characteristics. These include identifying conditions under which the detector fails, as well as those in which it works well.

One of the best-known anomaly detectors that has been applied to intrusion detection is stide[1]. Developed at the University of New Mexico, stide aims to detect attacks that exploit processes that run with root privileges. The original work on stide presented empirical results indicating that sequences of length six and above were required for effective intrusion detection.

This paper presents an evaluation framework that maps out stide's effective operating space, and identifies the conditions that contribute to detection strength, blindness or weakness. A theoretical justification for why sequence lengths six and above were effective is given, and the consequences of a different choice on detector performance is explained.

In addition, we give results of our investigation, which characterizes regions of the anomaly space in which stide is capable of anomaly detection and those in which it is not. We believe that relating detector properties of this kind to manifestations of intrusive activities is necessary if effective anomaly-based intrusion detection systems are to be built and deployed.

1 Introduction

In a solid body of work inspired by the way the natural immune system distinguishes *self* from *other*, Forrest et al. [3] presented and analyzed the effectiveness of a detection scheme aimed at enhancing the security of computer systems. Analogous to the natural immune system, computer system security was seen as an instance of the more general problem of distinguishing self, e.g., the normal behavior of system programs, from other, e.g., the behavior of trojanized system programs. The resulting anomaly detector was initially presented as a change-detection algorithm applied to the detection of computer viruses [2]. It has since been applied to the task of detecting intrusions or exploits by way of detecting abnormal behavior in processes that run with root privileges on UNIX systems. Named "stide", (Sequence TIme-Delay Embedding), the anomaly detector was designed to operate on categorical data in the form of system kernel calls issued by the running process to the kernel of the host system. The reference to "time" in the name of the detector reflects the time-series nature of the categorical data upon which the detector was deployed.

Through the series of papers that have documented the many experiments aimed at studying the effectiveness of stide with respect to the detection of exploits and intrusions in UNIX systems, the one curiosity that has been most conspicuous due to its significant impact on the performance of the detector, has been the question of the "best" or most appropriate detector-window length or sequence length required in any application of the algorithm. Note that in this study, the terms detector-window length (DW), sequence length and sequence size are used interchangeably to refer to the number of individual, categorical elements that make up a sequence. For stide, this value is set *a priori*, and used to determine the length of all the sequences obtained from both training and test data. In the literature, we find that a sequence length of six occurs consistently with stide in the experiments performed by the authors of the detector. For example, although a sequence length of 10 was finally settled on, in the results for the experiments in [5], it was also observed that a sequence length of *at least* six appeared to be necessary in order to detect all the intrusive data presented to the detector. This observation naturally prompts questions regarding the appropriate value of the detector-window parameter for stide, a problem that is not at all foreign to the community [12]. Questions such as:

- if not by "ad-hoc means" [12], how else can the "best" detector-window length determined?

- why does a detector-window length of six appear to work, and not lengths less than six?

- will it be the case that six will be appropriate for all data from differing environments?

- what is the impact on accuracy if an incorrect detector-window length is set? That is, what happens if six is selected but it is not the optimal length?

That the value of the sequence-length parameter impacts the performance of the detector has not only been noted by the original authors, but also in subsequent, independent, work [12, 9, 5]. Lee & Xiang [12] did propose an information theoretic solution to the problem of choosing the optimal detector-window length for stide. We address their solution in section 3.

The question of the appropriate detector-window length may have implications for aspects of detection other than performance. In particular, we were interested to know whether the results obtained by the original investigators represent a serendipitous match between the particular data sets used and the detector-window length. In other words, is six a necessary parameter value for this kind of detector, or simply sufficient for the data at hand? We feel that answering questions of this nature is essential if we wish to avoid deploying sensors of this kind (anomaly detectors in general, not just stide) in environments where they will simply fail.

Our study into the detection efficacies of stide have vastly increased our understanding, not only of the performance of anomaly detectors, but also of the anomaly-detection process itself, as it is applied to intrusion detection. We would like to note at this point that it would have been extremely difficult, if not impossible, to validate our research on the performance of anomaly detectors, particularly stide, if it were not for the generosity of the researchers at the University of New Mexico in making their datasets, detector and documentation readily available.

2 A brief description of stide

Stide acquires a model of normal behavior by segmenting the training data into fixed length sequences. This is done by sliding a detector-window of length DW over the training data. Each length DW sequence obtained from the data stream is stored in the "normal database" of sequences of length DW. A similarity metric is then used to establish the degree of similarity between the test data and the model of normal behavior obtained in the previous step. Again sequences of length DW are obtained from the test data using a sliding window, and for each length DW sequence, the similarity metric simply establishes whether that sequence exists or does not exist in the normal database. A length DW sequence from the test data that is found to exist in the normal database (where "existing" requires that an identical sequence be found in the normal database that matches the sequence obtained from the test data), is assigned the number 0. Sequences that do not exist in the normal database are assigned the number 1. The decision is binary, there is an exact match for a sequence from the test data in the normal database (0) or not (1).

The detector's final response to the test data, or anomaly signal, involves a parameter known as the "locality frame". The locality frame is a value determining the length of a temporally local region over which the number of mismatches are summed up. For example, if the locality frame is set to 20, then at each point of the test data the number of mismatches in the last 20 sequences including the current sequence is determined. The number of mismatches that occur within a locality frame is referred to as the locality frame count (LFC). The locality frame count is the final anomaly signal that is used to determine how anomalous the test data is. The length of the locality frame is a user-set parameter that is independent of the length of the detector-window used to segment both training and test data.

3 Conditional entropy and stide performance

In a paper presented in 2001 [12], it was suggested that the conditional entropy of the intrusive sequences was a key factor contributing to the optimality of six-symbol sequences in the stide data. Plots of the conditional entropy of the UNM `sendmail` data are shown in [12, Figure 1] where it is suggested that the knee in the data that appears in the plots indicates that little or no additional information occurs beyond the fifth or sixth symbol in the sequence.

In this experiment we aim to show that contrary to the suggestion in [12], conditional probabilities do not affect stide. In order to do this we need to establish pairs of training and test data that differ *only* in terms of increasing irregularity (measured as conditional entropy) and nothing else. This means that the alphabet size, alphabet symbols and sample size are all kept constant, while irregularity is calibrated to increase at fixed and steady intervals. We used 11 streams of training and test data pairs [6] that comply with these requirements. The data-generation process does not introduce anomalous sequences or symbols into the test-data stream. The reason for this is because introducing obviously-anomalous phenomena into the data stream would confound the results of the experiment; we would not know whether the detector was responding to the fluctuations in data regularity or to the presence of anomalous sequences.

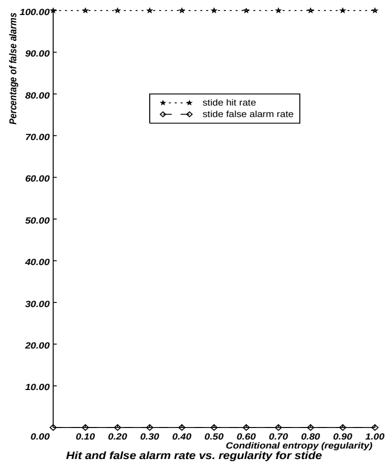

Figure 1: Hit and false alarm rate for stide

The data pairs are labelled 1 to 11, and each pair differs from the preceding pair in terms of a measured increase in irregularity. The training and test-data pair labelled 1 are therefore the most regular, and the pair labelled 11 are completely random data. For details on the generator and data-generation technique, see [6, Section 2]. Into each of these 11 datasets we inject a single anomaly consisting of a single symbol not present in the training data. This is the simplest unequivocally anomalous event that stide can be expected to detect. The detector-window length for stide in this experiment was set to 2 to be consistent with the data generator in which probability each element depended only on the value of the previous element in the sequence.

For stide, a hit occurs when a mismatch is registered. In the case of our test data, this will occur whenever the anomalous character is within the detector window. The effect of the locality frame was ignored, because the locality frame only serves to magnify or enhance the anomalies, i.e., mismatches, that may have clustered within a temporally local region. In our case there is only a single anomaly; it produces a single cluster consisting of two consecutive mismatches as the anomalous symbol passes through the detector window of length 2. An anomaly or mismatch anywhere else will be regarded as a false alarm regardless of whether or not it occurs in some temporally local region; therefore we focus only on whether or not the basic anomaly, in terms of a sequence mismatch, was registered.

Figure 1 presents experimental results in terms of hits and false alarms. We can see that, given a situation where everything was kept constant, including the type of anomalous phenomenon introduced into the data streams, stide remained unaffected by the regularity increase from one data stream to the next, and continued to detect the anomalous symbol present in each of the 11 test-data streams. These results appear sensible, because stide has no notion of probability, and will only be affected by probability if some aspect of probability introduces anomalous sequences into the test data. If the data-generation process does not introduce anomalous sequences, the fluctuations in data regularity itself, in isolation, makes no impact on the ability of stide to detect the anomalous symbol. If data regularity, measured as conditional entropy, does not affect the stide detector, then it is highly unlikely that this aspect of categorical data would be the determining factor for the appropriate sequence length that must be employed by stide.

4 Sequences: rare, common, and foreign

Stide characterizes the normal behavior of a monitored process in terms of a database comprised of sequences of length DW. These sequences are obtained by sliding a window of this length along the trace (or traces) of system calls that have been obtained from the monitored process in the absence of intrusions. Stide only checks to see if a test sequence is in the database or not, but the *frequency* with which each sequence occurs in the traces is key to the exposition in section 5.3 below.

For purposes of that discussion, we define a *rare* sequence as one that occurs infrequently in the training data. For our purpose, we arbitrarily define as rare, sequences that have a frequency of occurrence of less than 0.5% in the normal traces. All others are considered to be *common* sequences.

Foreign sequences are those that do not occur at all in trace(s) that were used to define normal behavior. Note that a sequence can be foreign by virtue of containing:

- foreign symbols, i.e., symbols that are not contained in the alphabet set of the training data, or
- a foreign order of symbols, i.e., a sequence in which each individual symbol within the sequence is a member of the training-set alphabet, but where the order of the symbols is one that does not exist in the set of sequences obtained from the training-set, or
- combinations of both.

In this work we focus specifically on the second condition, where a foreign sequence is foreign by virtue of the foreign order of its constituent symbols.

The term *minimal foreign sequence* is defined as a foreign sequence of the second type, having the property that all of its proper subsequences do exist in the trace(s). Put simply, a minimal foreign sequence is a foreign sequence that contains within it no smaller foreign sequences, [10].

4.1 Reproducing the original experiments

As a result of the present study, we are able to show that a detector-window of at least six was required to detect all intrusive traces in the Hofmeyr [5] dataset because of the existence of a very specific type of anomaly in the data which we describe as a minimal foreign sequence composed of rare or common subsequences.

In order to verify this, we first reproduce the experiment documented in Hofmeyr, et al [5]. It was this experiment that gave rise to the "why six" question. We do this using their Hamming-distance-based similarity metric, and then apply stide to the same data specifically to demonstrate that the "why six" question applies to the performance of stide, irrespective of the differing similarity metrics.

The data on which the experiment was run was obtained from the University of New Mexico website [4]. Table 1 provides summary information about the data.

The results presented in Figure 5 show the response of the detector that employed the Hamming-distance similarity measure. This graph only shows the curves associated with the sunsendmailcp, decode and syslogd attacks,

Program	Normal Data		Intrusion Data		
	No of traces	No of lines	Name of attack	No of traces	No of lines
Synthetic sendmail (UNM)	346		sunsendmailcp	3	1119
			decode	12	3067
			forwardingloops	10	2569
Synthetic sendmail (CERT)	294		syslogd	23	6504
			unsuccessful intrusion sm565a	3	275
			unsuccessful intrusion sm5x	8	1537
Synthetic ftp	8		wu.ftpd	5	1363
Synthetic lpr	9		lprcp	1001	164,232
Live lpr (MIT)	2698		lprcp	1001	165,248
Live lpr (UMN)	1231		lprcp	1001	164,232

Table 1: The data obtained from the University of New Mexico, and used to replicate the experiments documented in [5].

and was done to replicate the graph of results presented in [5]. We find in our results, as did they, that the important features are present in both cases, namely the absence of an anomaly signal for sequence lengths of less than six in the intrusive trace labelled "decode 1", which corresponds to the file named sm-280.int for the kernel calls associated with the PID 283.

As can be seen in both Figure 5 and Figure 6, the decode intrusion is not detectable for sequence lengths of less than six. The implication of these results, as stated in [5], is that a sequence length of six or greater is required because that will allow the detection of anomalies in all intrusive traces.

5 Effect of minimal foreign sequences on stide's performance

We have shown above that (ir)regularity in data, as measured by conditional entropy, does not affect the stide algorithm, nor does it determine the appropriate detector-window length to set, as stated by [12]. Consequently, we seek to identify the phenomenon that does determine the appropriate detector-window length.

We hypothesize that a detector-window length of at least six is required to detect all intrusive traces in the Hofmeyr experiments, because the length of the smallest minimal foreign sequence present in one of the intrusive traces was six, and that this minimal foreign sequence of length six must have been composed of either rare or common subsequences. This explains why stide or the Hamming distance detector did not detect this anomaly for detector-window lengths of less than six, i.e., because the smaller subsequences that make up the minimal foreign sequence of length six already exist in the training trace that produced the normal database.

The strength of the stide algorithm is in the detection of foreign sequences. The stide algorithm is capable of detecting foreign sequences simply because foreign sequences match no other sequence in the normal database. However, factors such as the relation between the length of the sliding window and the length of the foreign sequence, as well as the effect of sliding a window over the foreign sequence, do make a significant impact on the detection capabilities of stide. The following experiments will show that it is possible for a foreign sequence to be composed of a mixture of rare and common sequences, and that the length of the smallest minimal foreign sequence in a given trace is what prescribes the appropriate detector-window length for the stide detector.

We will use a detector based on Markov models [6] as a tool for comparison in order to help illustrate the factors that do or do not affect stide, but that may or may not affect another detector employing a different approach. We will refer to the detector based on Markov models as the Markov detector. The Markov detector employs conditional probabilities in its function as an anomaly detector. Briefly, it determines the probability of seeing an event, given the previous N events. The following is an outline of the experimental procedure that we will employ to show that the length of the minimal foreign sequence prescribes the length of the appropriate detector window that must be set to detect anomalies in a stream of data known to contain the manifestations of an intrusion, fault or attack.

- Generate training data;
- Generate background test-data stream;
- Use the training data to select minimal foreign sequences of lengths 2 to 9, composed of rare subsequences;
- Inject the anomalies into the generated background test-data stream to create the final stream of test data;

- Deploy both anomaly detectors (stide and Markov) on the same training and test data, while varying their detector-window lengths with respect to the length of the injected anomalous sequence;

- Record the response of detector to the injected anomaly.

5.1 Constructing training data

The training data were constructed using a Markov-model transition matrix. The precise method for generating the training data is documented in [6]. Although numbers were used to represent the elements of the training-data stream, the numbers were treated as categories.

The transition matrix used to generate the training data had a conditional entropy value of 0.1. This means that at each point in the data stream, the next element is highly predictable given the current element, i.e., there is low uncertainty as to what the next element will be. Such a transition matrix was chosen simply because it generated data with the following characteristics:

- A large proportion of the data consists of a repetition of the sequence 1, 2, 3, 4, 5, 6, 7, 8. Ninety-eight perecnt of a one-million-element data stream, generated with this transition matrix, will consist of a repetition of the sequence 1, 2, 3, 4, 5, 6, 7, 8. This results in a consistent set of obviously common sequences, regardless of the length required of the sequences. This is particularly necessary when constructing the test-data stream. A test-data stream made up of commonly-occurring sequences is desirable in order to allow us to observe the response of a detector to the injected anomaly without being confounded by naturally-occurring rare or foreign sequences.

- Despite the repetition in a large portion of the data resulting in a usable set of common sequences, there is a small amount of unpredictability in the probabilities that populate the matrix which ensures the occurrence of rare sequences necessary for selecting the constituent rare subsequences in a minimal foreign sequence.

The alphabet size for the training data was 8. It is noted that alphabet sizes in real-world data are certainly much higher than this; for example, there are about 243 unique kernel calls in BSM audit data. However, the method aims to evaluate the capabilities of the detector in detecting the higher-level concept of an anomaly. Although alphabet size may play a role with respect to certain aspects of the data, such as influencing the size of the set of possible foreign sequences or the size of the set of possible sequences that populate the normal database, a foreign sequence is still a foreign sequence regardless of the alphabet size, and the concept of a rare sequence will also remain immutable regardless of alphabet size. This abstraction allows us to study the response of the detector using synthetic data, as well as to apply the results from the synthetic environment to real-world environment.

The aforementioned matrix was used to generate a training-data stream of 1,000,000 elements. The sample size of 1,000,000 was an arbitrary choice, selected so that the data set would not be insufficiently small. There were two parameters that were chosen arbitrarily in this experiment, the sample size of 1,000,000 elements, and the length of the minimal foreign sequence (AS), which ranged from 2 to 9.

5.2 Constructing background test data

The background data for the test-data stream consisted of the most commonly occurring sequences only, which given the training data described above, consists of a repetition of the sequence 1, 2, 3, 4, 5, 6, 7, 8. This ensured that only common sequences populated the background data. This was a desired property primarily because our aim was to observe the response of a detector to the specific minimal foreign sequence that we were going to introduce into the background data in the second phase of this procedure. We therefore wanted background data that would not interfere with a detector's response by containing within it any obviously anomalous event that may constitute noise to a particular detector, for example naturally occurring rare or foreign sequences.

5.2.1 Characteristics of the anomaly

We will be introducing an anomaly that consists of a minimal foreign sequence of length AS, composed of rare subsequences, into the data stream. As noted above, a rare sequence is defined to be a sequence that occurs less than 0.5% of the time in the training data.

The selection of rare sequences was prompted by the expectation that the Markov detector will have the ability to detect rare sequences. In cases where the length of the detector-window is decidedly less than the length of the anomaly AS, we encounter the situation where the detector does not "see" all of the minimal foreign sequence at once. Instead, the detector is relegated to producing an anomaly signal based only on the smaller subsequences that pervade the larger minimal foreign sequence. Under such situations, we would like to observe the effect of the rare subsequences on the performance of both the Markov-based detector and stide. Although we already know that stide does not have the ability to respond to rare sequences, we will nevertheless apply the stide detector to an anomaly with these characteristics, primarily for the sake of charting and comparing the performance space of both detectors in an attempt to quantify how much more the ability to detect rare sequences actually confers upon the detection of foreign sequences under such circumstances.

5.3 Producing and injecting minimal foreign sequences

The minimal foreign sequences and their constituent subsequences must now be carefully chosen so that the injection process itself does not introduce unintended perturbations in the background data. This is particularly significant with respect to the sequences at the boundaries, i.e., where some elements of the injected anomalous sequence and some elements of the background data may combine within a detector's window to produce sequences that affect the anomaly detector in uncontrolled and unintended ways. In particular, we want to avoid producing additional, undesired, foreign sequences due to the combination of symbols from the injected sequence and surrounding symbols from the trace.

We have determined that 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 foreignness and minimality. These same properties complicate the problem of injecting the anomaly, which remains somewhat of an art. Essentially, the problem is one of ensuring that all of the $2(\mathrm{DW}-1)$ sequences of length DW that can be composed at the boundary of the injection, using contiguous symbols from the anomaly and the background trace, are actually in the database. If this is not the case for some location in the trace, a new anomaly must be produced 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 repeated for each detector-window length of 2 to 15. Note that the length of the detector window dictates the length of the subsequences that compose each minimal foreign sequence. In total we have 112 test data streams.

5.4 Deploying detectors

We deployed the stide and Markov based detectors on the suite of data created in the preceding sections. For each minimal foreign sequence being detected, we varied the length of the detector window from 2 to 15. It should be noted that for stide we ignored the locality frame count, focusing on the indication of a match (0) or mismatch (1). We reasoned that although further processing can be performed on the results of the similarity measure for purposes of smoothing away noise or enhancing signal strength, no amount of subsequent processing can compensate for the underlying inability to detect a specific phenomenon.

The locality frame count (LFC) sums up the number of mismatches experienced within the span of the locality frame. Although the LFC does contribute to the final anomaly signal, it only comes into play *after* a sequence has been determined to be a match or mismatch. If the detection of a foreign sequence is missed, meaning that it does not register as a mismatch, then no amount of applying the LFC or adjusting its length will cause the missed anomaly to be detected.

5.5 What is meant by hit, miss, detection blindness and detection weakness?

The results of our experiments are expressed in terms of hits and misses, and in terms of regions of detection blindness and weakness. When a detector window slides over an anomaly, e.g., a foreign sequence, at various points of its journey it will view sequences that are composed of a combination of the elements from the foreign sequence and elements from the background data. Under such circumstances, the interaction between the elements of the foreign sequence and the background data will cause sequence types to arise that prompt the anomaly detector to respond in one fashion or another. Regardless of how the detector responds, the response is still influenced by elements of the foreign sequence. Only when the detector window completely clears the entire foreign sequence (i.e., no elements within the detector window belong to the foreign sequence), can we say that the response of the detector is no longer influenced to the foreign sequence. In other words, as long as some part of the foreign sequence is viewed by the sliding detector window, it can be argued that the detector's response is due to the presence of the foreign sequence in the data. As a result, the response of the detector in such a circumstance should also be considered in the process of determining hits or misses. This line of reasoning resulted in the concept of the incident span that we use to determine hits and misses. The incident span includes the DW $-$ 1 elements of the background data adjacent to the anomalous sequence on one side of the detector window, the AS elements of the anomalous sequence, and the $DW - 1$ events of the background data adjacent to the anomalous sequence on the other side of the detector window (see Figure 2). The length of this span is therefore $AS+2(DW-1)$ elements. Alternatively, it can be said that $AS + (DW - 1)$ sequences of length DW are contained within the incident span.

Using the situation where only a single anomaly was introduced into each test stream, and letting the detector response range from 0 (indicating completely normal) to 1 (indicating maximal abnormality), we describe a detector 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 that is not definitely normal has been seen;
- 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 Markov 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.

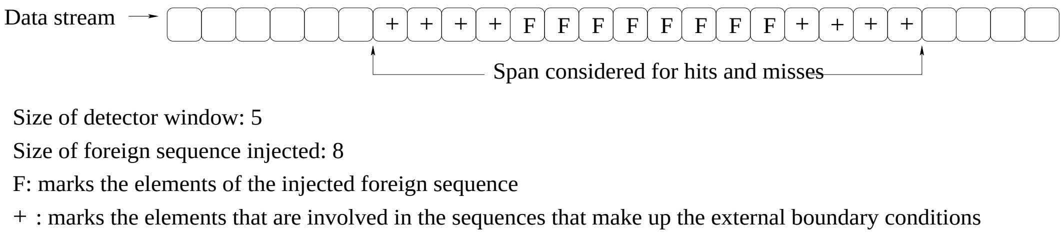

Figure 2: The incident span. A detector's response to the sequences in the incident span are considered when determining hits and misses.

To avoid the compounding effects of varying the threshold of the Markov detector, we set its threshold at 1, recognizing only maximally anomalous (foreign) sequences as "hits."[1]

5.6 Results and discussion

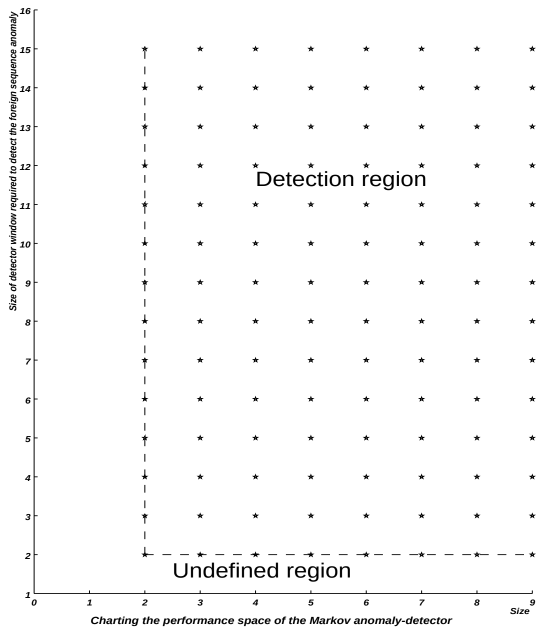

Figure 3: Markov detector efficacy

Figures 3 and 4 show the results of the experiment presented above. They map the detection capability of both the Markov detector and stide with respect to an injected foreign sequence composed of rare sequences.

The x-axis marks the increasing length of the minimal foreign sequence injected into the test-data stream, and the y-axis charts the length of the detector window required to detect a minimal foreign sequence of a given length. Each star marks the length of the detector window required to detect a foreign sequence whose corresponding length is marked on the x-axis, where the term "detect" specifically means that a maximum anomalous response occurred in the incident span. The areas that are absent of a star indicate that the detector was

[1] Detection thresholds are often used to determine "alarm-worthy" events. The most anomalous detector 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.

unable to detect the foreign sequence whose corresponding length is marked on the x-axis, where unable to detect means that the maximum anomalous response recorded along the entire incident span was 0, signifying completely normal.

Since the Markov detector is based on 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 3 begins at 2. Although it is possible to run stide 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 on the side of the Markov detector, argued against running stide with a window of 1.

The x-axis also begins at 2. 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 length 1, 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, both Figures 3 and 4 show an undefined region corresponding to the detector-window and anomaly length of 1.

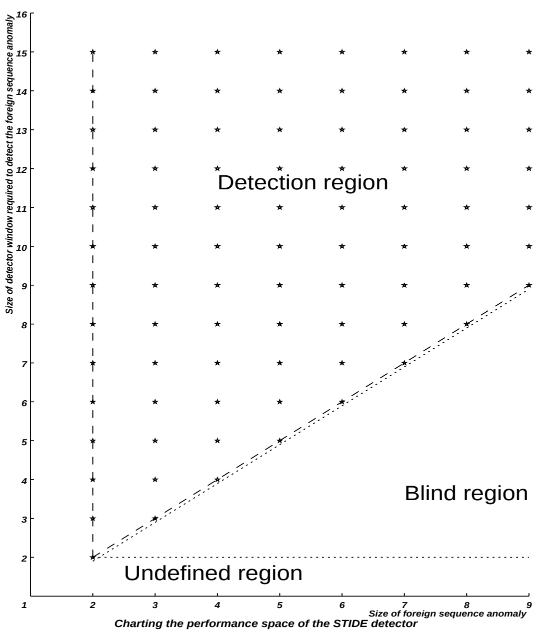

Figure 4: Stide detector efficacy

The results show that although the stide and Markov-based detectors both use the concept of a sliding window, and are both expected to be able to detect foreign sequences, their differing similarity metrics significantly affect their detection capabilities. There are three main points to note from the results. First, for stide, the detector-window length parameter must be greater than or equal to the length of the foreign sequence. The minimum length of the detector window required to detect each minimal foreign sequence is the size of the minimal foreign sequence itself. As can be seen from diagonal line in the results, the correlation between detector-window length and anomaly length is strong: $y = x$. Second, the results show that the similarity metric used by each detector significantly affects detection performance. In stide's case, even though we know 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 similarity measure employed by stide appears to have a weakness in that it is unable to detect minimal foreign sequences composed of rare subsequences under conditions where $DW < AS$. As a result, there are no guarantees that stide will detect faults even if they do manifest as foreign sequences in the data. The Markov 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 suggests that there are factors in *this* data stream that favor detectors that employ conditional probabilities. These factors, however, appear to have no effect on the sequence-matching approach employed by stide. Finally, by charting the performance of stide and the Markov detector with respect to the detection of minimal foreign sequences, we are able to observe the nature of the gain achieved in detection performance between an algorithm that employs conditional probabilities and one that employs the sequence-matching scheme used by stide. This gain in detection ability, due to the use of conditional probabilities, is significant and is illustrated by the blind region marked out in Figure 4.

These results provide evidence that shows a strong relationship between the length of the minimal foreign sequence and the length of the detector window required to detect such a phenomenon. It appears that the appropriate sequence length for stide is prescribed by the length and composition of the minimal foreign sequences present in the data.

6 Locating minimal foreign sequences in real-world data.

In the previous section we saw that some phenomena in the "decode 1" intrusive trace caused both stide and the Hamming-distance-based detector to

	dec.1	dec.2	snsndmailcp	flps-1.162	flps-1.163	flps-2.170	flps-3.182	flps-3.183	flps-4.206	flps-4.207	flps-5.119
1											
2		2	10	7	4	3	8	3	6	4	3
3		1	13	7	2	1	8	1	8	2	1
4			4	1	4	2		2	1	4	
5			2	2			2		2		
6	1	1									
7				2		2	1		2		
8						1					
9			1	1			1		1		
10											
11				1			1		1		
12											
13											
14											
15											
16											
17											
18											
19											
20							1				

Table 2: The raw number of minimal foreign sequences of lengths 1 to 20 for each named intrusive trace. The empty cells mean that no minimal foreign sequences of that length could be found in the trace. Note the single length-six minimal foreign sequence in the dec.1 (decode 1) column. The smallest minimal foreign sequence in every other trace is of length 2. In order for stide to detect all intrusive traces, a detector window of length 6 is required.

completely miss the anomalies present in the "decode 1" intrusive trace when detector-window lengths of less than six were employed. In experiments with synthetic data, we found that such behavior is typical of both detectors in the presence of minimal foreign sequences composed of rare or common subsequences. This final section ties these observations together, and proposes that the solution to the "why six" problem lies in the presence of a length-six minimal foreign sequence, composed of rare or common subsequences, in the decode 1 intrusive trace. Since no minimal foreign sequences exist in the decode 1 trace with lengths less than six, unlike all the other intrusive traces, no anomalies could be detected when detector-window lengths of less than six were used. This meant that a detector-window length of six was necessary in order to detect anomalies in *all* intrusive traces, including decode 1.

Our task at this point is to identify the minimal foreign sequences that are present in the Hofmeyr data. We wish to chart characteristics, such as their constituent subsequences, and their various lengths.

This serves three purposes:

- to show that the anomaly types we laid out in the anomaly domain actually exist in real-world data;

- to show that regardless of the data, i.e., synthetic or real-world, when the performance of a detector has been established with respect to the anomaly types described in [7], the performance results for a detector are immutable, and will persist reliably across datasets;
- to verify that, in the case of these detectors, it is the presence of these anomaly types in the data stream that dictates the appropriate detector-window length to set;
- to solve the "why six" problem.

We proceed to identify all the minimal foreign sequences of lengths 2 to 20 in the sunsendmailcp, decode and syslogd intrusive data, using all the normal data for the system programs associated with those intrusive traces. We concentrate only on these traces, because the observation made by Hofmeyr et al. regarding the fact that a sequence length of at least six was required to detect anomalies in all the intrusive traces, was made on these traces from the detection results presented in [5].

Table 2 lists the length and number of minimal foreign sequences present in the intrusive traces decode 1, decode 2, sunsendmailcp and forwardingloops. We can see from the table that decode 1 contains only one minimal foreign sequence of length six, whereas in every other intrusive trace the smallest minimal foreign sequence was of length 2. This means that stide required a detector-window length of six in order to detect that single anomaly in decode 1 because, there were no minimal foreign sequence anomalies of lengths *less* than six to detect in that intrusive trace. Upon further analysis of the single minimal foreign sequence in decode 1, we find that it is actually a minimal foreign sequence of length 6 with rare subsequences. Precisely:

```
Filename: sm-280.int283.

Actual Sequence: 2, 95, 6, 6, 95, 5

Translated to system calls: fork, connect, close, close, connect, open

Start line number: 79

End line number:84
```

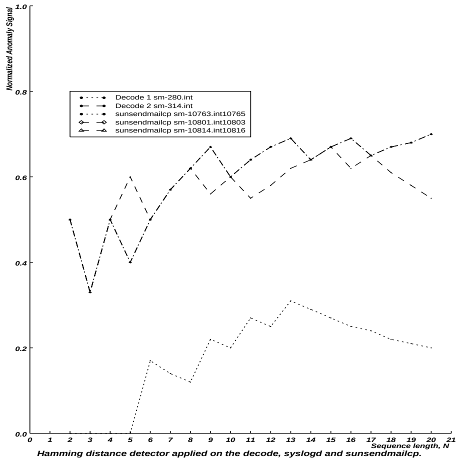

Figure 5: Normalized Hamming-distance similarity measure plotted against sequence length DW for a detector employing the Hamming-distance similarity measure. We replicated the experiment documented in [5] to validate the observation that a detector-window length of at least six is required to detect all intrusive traces.

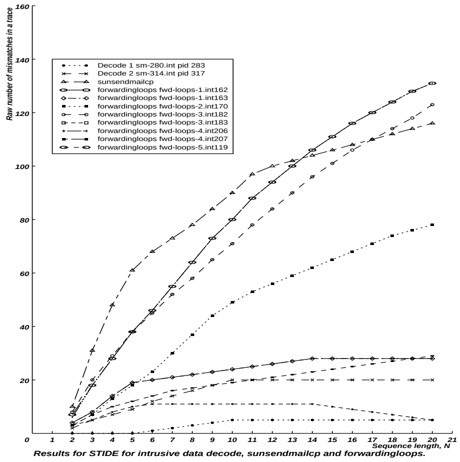

Figure 6: The response of stide plotted against sequence length. The same observation made in [5] can be made here as well. It appears that at least a sequence length of six is required to detect anomalies in *every* intrusive trace.

7 Conclusions and future work

From the series of experiments above, we have confirmed our hypothesis that a detector-window length of at least six was required to detect all intrusive traces in the experiments in [5]. This was because the length of the smallest minimal foreign sequence present in one of the intrusive traces was six. We found that the intrusive trace labelled "decode 1" contained a single size-six minimal foreign sequence, composed of rare subsequences. The rare subsequences meant that only when the detector-window length was large enough to see the entire minimal foreign sequence would that sequence register as an anomaly.

Detection accuracy will be compromised in situations where the length of the detector window employed by stide is set to be smaller than the length of the smallest minimal foreign sequence. In such cases, attacks that may manifest as those minimal foreign sequences will elude the detector altogether. Since $DW < AS$, no anomalies will register in the data stream. However, if a larger detector-window length is used, the minimal foreign sequences will suddenly be detectable.

We showed the effect of minimal foreign sequences composed of rare or common subsequences on stide's performance, and how their presence undermines the claim that stide will detect foreign or "unusual" sequences that occur in a stream of data. We have identified the conditions under which stide is completely unable to detect the presence of foreign sequences in a data stream. Identifying minimal foreign sequences, and establishing their effect on stide, enabled us to provide a solution to the question of the "best" or most appropriate detector-window size to select in any application of the stide algorithm.

We have also shown that the performance characteristics established for stide on synthetic data remained pertinent across datasets. In this case, even when the detector was deployed on real-world data, we were able to explain its performance behavior using the lessons learnt for that detector on synthetic data.

As a final note, we remind the reader that we are assuming that the foreign sequences we encountered in the real-world data actually are the manifestations of the intrusions of interest. We are currently not aware of any analysis that established whether the intrusions actually manifested in the system call data obtained from the "strace" sensor, and whether those manifestations were anomalous. As a result, we really do not know if the foreign sequences we identified in the intrusive data were or were not the result of the intrusions. This makes it hard to determine if the detection of those foreign sequences were hits or false alarms. It is equally likely that the single minimal foreign sequence of size six in the decode 1 trace was the result of insufficient training data.

These speculations raise a more general issue. To what extent can we establish a link between detectable anomalies and intrusive behaviors? How can we decide, *a priori*, what kind of a sensor stream is appropriate and what detector characteristics are likely to be well matched to the stream. For example, the "decode–1" intrusion is characterized, in the UNM data, by exactly one minimal foreign sequence of length six. We have shown that stide, with a window size of less than six, cannot detect this particular incident. Are there intrusive scenarios that would produce minimal foreign sequences with greater lengths? In a similar vein, given knowledge of the detector and the working definition of normal, i.e., the database, is it possible to either modify an attack so that its trace appears to contain only normal sequences, or so that it contains only minimal foreign sequences of length greater than the size of the detector window? We are beginning to investigate these questions, and preliminary results indicate that escaping detection in these ways is possible for stide-like detectors. We would like to extend these investigations to other anomaly-detection schemes.

8 Acknowledgements

The work herein was supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under contracts F30602-99-2-0537 and F30602-00-2-0528. Many other people also contributed in various ways; the authors are grateful to Kevin Killourhy, Sami Saydjari and Tahlia Townsend for their help. The authors particularly wish to thank John McHugh for his time and contribution to this paper. This paper draws on Kymie Tan's forthcoming dissertation [10].

References

[1] Stephanie Forrest, Alan S. Perelson, Lawrence Allen, and Rajesh Cherukuri, "Self-nonself discrimination in a computer", In *IEEE Sympo-*

sium on Research in Security and Privacy, pp. 202-212, IEEE Computer Security Press, Los Alamitos, CA, 16-18 May 1994, Oakland, CA.

[2] Patrick D'haeseleer, Stephanie Forrest and Paul Helman, "An immunological approach to change detection: algorithms, analysis and implications", *Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy,* pp. 110-119, IEEE Computer Society Press, Los Alamitos, CA, May 1996, Oakland, CA.

[3] Stephanie Forrest, Steven A. Hofmeyr, Anil Somayaji, and Thomas A. Longstaff, "A sense of self for Unix processes", In *Proceedings 1996 IEEE Symposium on Security and Privacy,* pp. 120-128, IEEE Computer Society Press, Los Alamitos, CA, May 1996, Oakland, CA.

[4] University of New Mexico, "Computer Immune Systems", Internet: http://www.cs.unm.edu/ immsec/data-sets.htm, 2000.

[5] Steven A. Hofmeyr, Stephanie Forrest, and Anil Somayaji, "Intrusion detection using sequences of system calls", *Journal of Computer Security,* vol. 6, no. 3, pp. 151-180, 1998.

[6] Roy A. Maxion and Kymie M. C. Tan, "Benchmarking anomaly-based detection systems" *International Conference on Dependable Systems and Networks*, Los Alamitos, California, 2001, pp. 623-630, IEEE Computer Society Press, 25-28 June, New York, New York.

[7] Author-1 and Author-2, "Removed for purposes of blind review".

[8] Author-1 and Author-2, "Removed for purposes of blind review".

[9] M. Stillerman, C. Marceau, and M. Stillman, "Intrusion detection for distributed applications" *Communications of the ACM*, 42(7), pp. 62-69, July 1999.

[10] Author-1, "Removed for purposes of blind review".

[11] Christina Warrender, Stephanie Forrest and Barak Pearlmutter, "Detecting intrusions using system calls: Alternative data models", In *Proceedings of the 1999 IEEE Symposium on Security and Privacy,* pp. 133-145, IEEE Computer Society Press, Los Alamitos, CA, 9-12 May 1999, Oakland, CA.

[12] Wenke Lee and Dong Xiang, "Information-theoretic measures for anomaly detection" In *Proceedings of the 2001 IEEE Symposium on Research in Security and Privacy*, pp. 130-134, IEEE Computer Society Press, Los Alamitos, CA, May 2001, Oakland, CA.