RZ 3011 (# 93057) 03/30/98
Computer Science/Mathematics 13 pages

Research Report

Anomaly Detection for Markov Models

Mehdi Nassehi

IBM Research Division
Zurich Research Laboratory
8803 Rüschlikon
Switzerland

Research Division
Almaden · Austin · Beijing · Haifa · T.J. Watson · Tokyo · Zurich

Anomaly Detection for Markov Models

Mehdi Nassehi

IBM Research Division, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland, mmn@zurich.ibm.com

Abstract

An anomaly-detection scheme is presented here in which the monitored process is modeled as a Markov chain. It generates an alarm depending on the likelihoods that the monitored process exhibits intrusive behavior or normal behavior. The scheme maximizes the true alarm rate such the fraction of alarms which are false does not exceed a given limit. The false alarm rate can be computed based on a model of normal behavior, which can be extracted from a sample of normal behavior. The true alarm rate, on the other hand, requires a model of intrusive behavior, even though samples of intrusive behavior are not widely available. Instead, a novel game-theoretical approach is used to model intrusive behavior also based on the sample of normal behavior. The uncertainties in both models are approximated by normal distribution, which simplifies the analysis and provides more insight.

1 Introduction

Anomaly detection is an intrusion-detection technique in which any significant deviation of the current characteristics of a process from its expected characteristics generates an alarm. (For a survey of intrusion-detection systems, including anomaly detection, see [1].) The expected characteristics of the process are captured in a model, referred to as the model of normal behavior. Optimum anomaly detection also requires a model of intrusive behavior for estimating the true alarm rates. Each model is instantiated from a model class, which defines the anomaly-detection scheme. For example, if the monitored process is voice, the models would be continuous-state continuous-time; if the process is the sequence of system calls generated by a program [2], the models would be discrete-event discrete-time. By redefining the event alphabet, many discrete-event discrete-time processes can be modeled as Markov chains, where the future states of the process depend on the past states only through the current state. The contribution of this paper is an optimum detection scheme for Markov-chain models.

The scheme proposed here is based on the modeling technique developed for the rule-based anomaly-detection scheme presented in [3]. Nevertheless, these schemes are complementary for the following twofold reason. First, every rule-based model can be mapped to a Markov model and vice versa, where there is one-to-one mapping between the left-hand sides of the rules and the states of the Markov model. Second, in the rule-based anomaly-detection scheme the triggering of certain rules or, equivalently, certain state transitions, result in an alarm; that is, rules or transitions are tested individually, albeit in parallel. In the scheme presented here, transitions are monitored in an aggregate fashion. More specifically, the average frequencies of transitions over a window of states is maintained and observed for any anomalies with respect to the normal average frequencies. Therefore the scheme presented in [3] is a local scheme and the one presented here is an aggregate one.

The paper is organized as follows. In Section 2, we show the conceptual framework of the proposed anomaly-detection scheme based on the modeling of normal and intrusive behavior and the optimum detection algorithm. In Sections 3 and 4, respectively, we present the analysis underlying the conceptual framework and apply it to a specific example. The results are summarized in Section 5.

2 Conceptual Framework

Any anomaly-detection scheme has two components: a behavior characterizer and a detector, as shown in Fig. 1. The behavior characterizer takes as input a set of configuration parameters and a sample of normal behavior and generates a set of detection parameters, which are then stored on a disk. These parameters are loaded into the detector as it is initialized. It can then detect any anomaly, with respect to these parameters, in the process being monitored and generate an alarm.

Compared to the general scheme depicted in Fig. 1, the anomaly-detection scheme presented here is more specific in two ways. First, the process monitored is modeled as a Markov process that consists of a sequence of states. At any point in time the next state depends on the past states only through the current state; that is the past states provide no information about the next state in addition to that provided by the current state. For example consider the sample process shown in Fig. 2. The process can be modeled as a Markov chain represented by the state diagram in Fig. 3. The diagram consists of one node for each state and a directed arc for each state transition.

The second way in which the anomaly-detection scheme presented here is more specific than

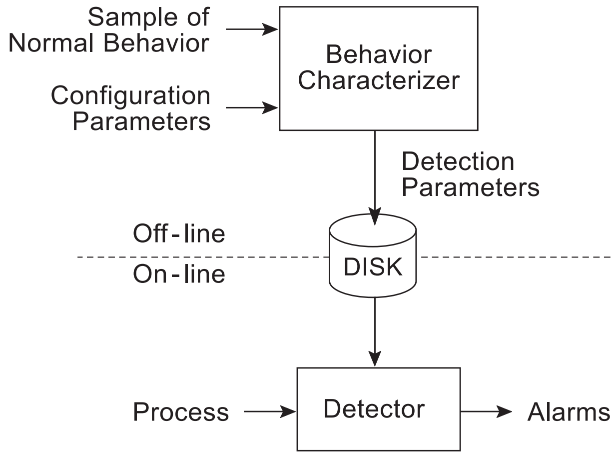

Figure 1: Anomaly detection.

$$S_1\ S_2\ S_3\ S_4\ S_2\ S_3\ S_2\ S_3\ S_4\ S_1\ S_1\ S_1\ S_2\ S_3\ S_4\ S_1$$

Figure 2: Sample of normal behavior.

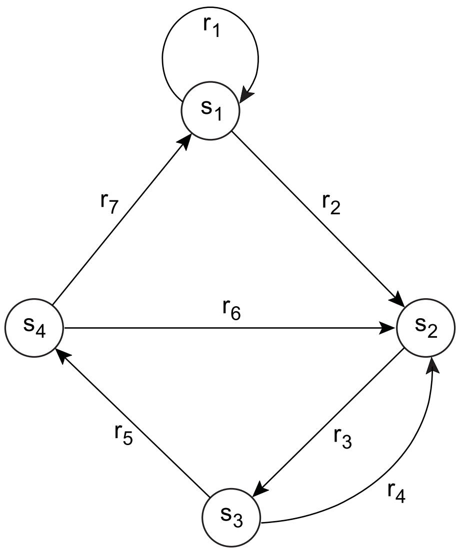

Figure 3: State diagram of normal behavior.

a general scheme is that the detection parameters are selected such that some measure of the detector's performance is optimized. In anomaly detection, two performance measures are of significance: the true and false alarm rates. The true (false) alarm rate is the mean number of successful (erroneous) anomaly detections per unit time. We formulate the optimization problem as the search for a set of detection parameters that maximizes the true alarm rate under the constraint that the ratio of the false alarm rate to the true alarm rate does not exceed a given value.

In order to solve the optimization problem analytically rather than empirically, the analysis such as the one presented in the next section must handle the mean true and false alarm rates, which requires a model of intrusive behavior and a model of normal behavior, respectively. Consequently, as shown in Fig. 4, the required behavior characterizer component consists

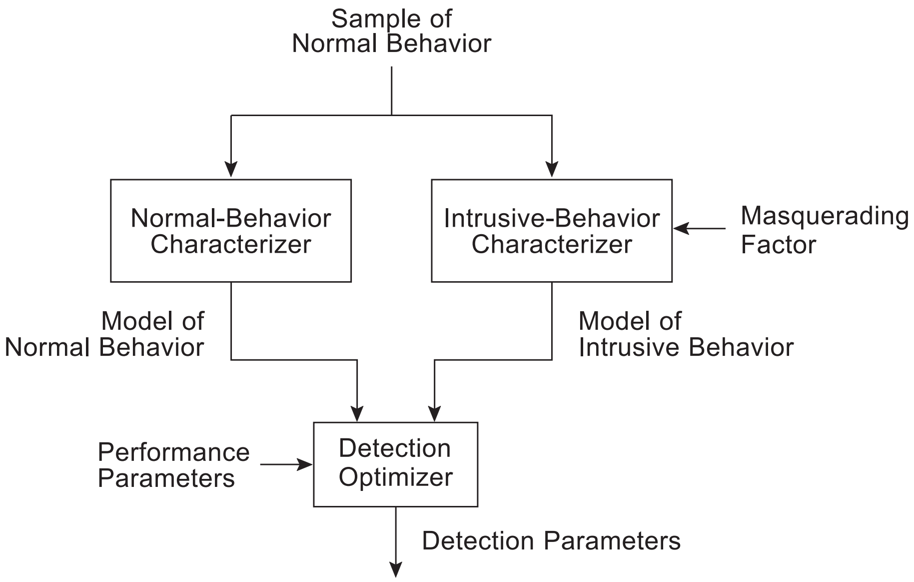

Figure 4: Behavior characterizer.

of three blocks: a normal-behavior characterizer, an intrusive-behavior characterizer, and a detection optimizer. Note that the model of intrusive behavior is generated from the sample of normal behavior and a parameter referred to as the masquerading factor. These blocks are described in the following subsections.

2.1 Modeling Normal Behavior

The model of normal behavior is extracted from a sample of normal behavior. This model allows us to predict the process for the case of no intrusions. We seek the best model in a class of models. This class is selected such that it abstracts those characteristics of the process that are of interest. Each model is an instantiation of the class for a set of parameters. Modeling the process under no intrusion means finding the model that best explains the sample of normal behavior.

Creating the state diagram is the first part of modeling. Consider the sample of a normal behavior shown in Fig. 2. This sample consists of a sequence of 16 states out of an alphabet of four distinct states, s_1 to s_4. The state diagram in Fig. 3 is created by representing each of these states by a node. A directed arc is drawn for state s_i to s_j if and only if $s_i s_j$ occurs in the sample.

The next stage in modeling is to assign frequencies to the state transitions. By frequency of a transition we mean the number of times a transition occurs in a sequence of states over the total number of transitions in that sequence. Note that using the frequency of transitions results in a more accurate model of the process than using the frequency of states because the former uniquely defines the latter but not the other way around.

The last stage of modeling is to identify the uncertainties in the frequencies of the transitions. Let K_D denote the number of transitions in the sample of normal behavior. In the next section we show that the uncertainty in the transition frequencies can be approximated

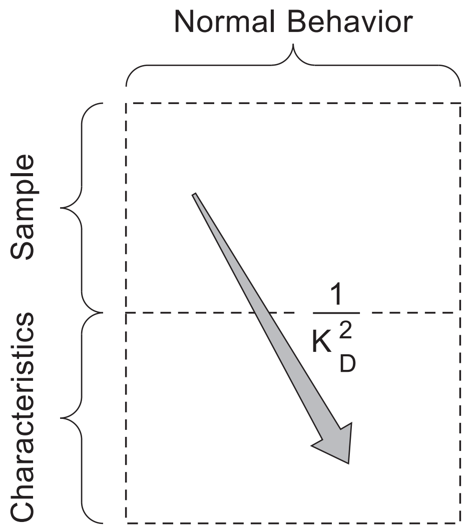

Figure 5: Uncertainty in the characteristics of the normal behavior from the detector's point of view.

by a normal distribution with a variance proportional to $1/K_D^2$. In Fig. 5 the increase of uncertainty from that of the sample, i.e. zero, to that of the model is represented by a labeled arrow.

2.2 Modeling Intrusive Behavior

Modeling intrusive behavior differs from modeling normal behavior due to the lack of extensive samples of intrusive behavior. In [3] a novel approach was developed based on the key idea of explicitly modeling the active role that the intruder plays in generating the intrusive behavior. Every intrusive behavior originates from an intruder whose objective is twofold: to carry out an unauthorized action and avoid or postpone being detected. The objective of the detector is also twofold: to advance the detection of intrusive behavior while reducing the probability of false detection. Thus anomaly detection is a game with two players: the detector and the intruder. Each player seeks an optimum strategy for maximizing its objective, whose actual value also depends on the strategy of the other player. In such a competitive game, one or more stable strategy pairs may exist [4]. However, except for very simple models of intruder and detector, finding these strategies is analytically complex.

To make the anomaly-detection game simpler, we decouple the intruder's strategy from the detector's. This decoupling can be achieved by assuming that the intruder subordinates unauthorized action to evasion from detection. Clearly, if the intruder gives priority to not being detected, it hurts the objective of the detector the most. Therefore, these assumptions help the intruder and, in effect, result in a model of the worst-case intrusive behavior. The optimum detection based on this worst-case model maximizes the lower bound on performance.

Giving evasion from detection priority, the intruder can identify the following optimum strategy: in order to delay detection, it will emulate normal behavior. It can emulate normal behavior based on whatever knowledge it has thereof. To quantify this knowledge, we assume that the intruder has a sample of the normal behavior of length K_I transitions. Like the detector, it can extract from this sample a model of normal behavior, as depicted in Fig. 6. The variance of uncertainty in this model is proportional to $1/K_I^2$. As the intruder emulates normal behavior, the characteristics of intrusive behavior are identical, from its point of view,

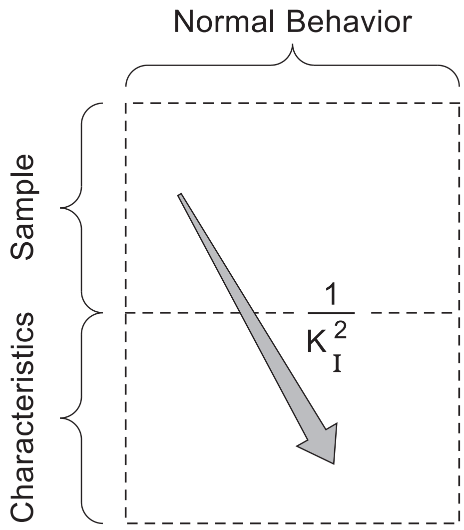

Figure 6: Uncertainty in the characteristics of the normal behavior from the intruder's point of view.

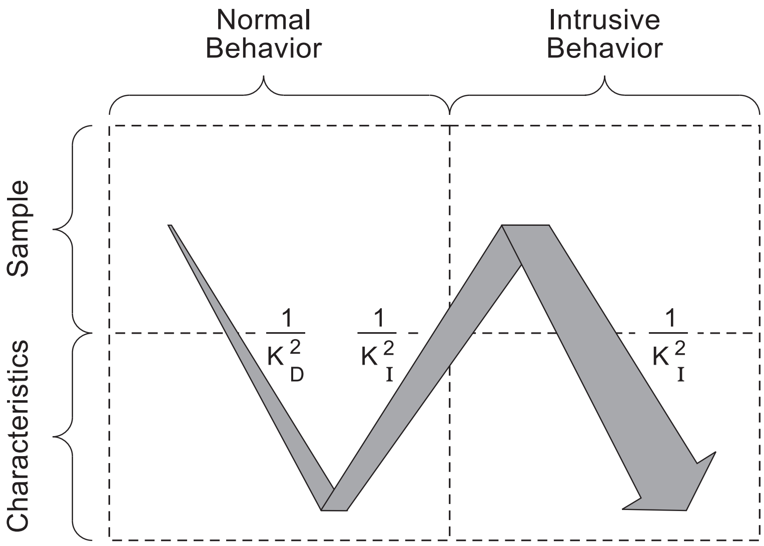

Figure 7: Uncertainty in the characteristics of the intrusive behavior from the detector's point of view.

to those of normal behavior. Here, we define the characteristics of a process given a sample as the distribution of the models of that process given the sample.

The characteristics of intrusive behavior from the detector's point of view can be obtained based on the characteristics from the intruder's point of view by accounting for the detector's uncertainty concerning the sample that the intruder possesses. The increase in uncertainty at each stage of the modeling is represented in Fig. 7. The resulting characteristics are those of intrusive behavior that we have been seeking. Note that it depends only on the sample of normal behavior and the parameter K_I. We refer to K_I as the masquerading factor.

The approach to modeling intrusive behavior presented here has the following features: (1) given a masquerading factor, a worst-case model of intrusive behavior can be generated; (2) the

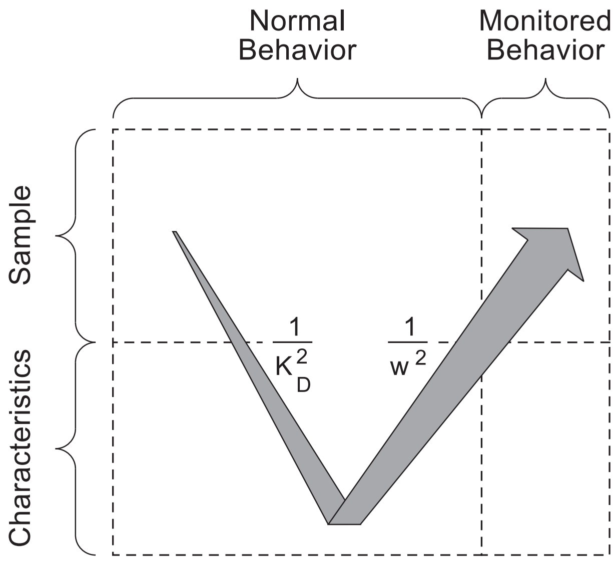

Figure 8: Uncertainty in the monitored sample of normal behavior.

factor is expressed in terms of the size of a sample of normal behavior, not the characteristics of normal behavior; therefore, it is not derived by a transformation of the characteristics of the normal behavior but rather directly from the sample of normal behavior; (3) for a factor of 0, the characteristics of intrusive behavior take their *a priori* value. As the factor is increased, the characteristics approach those of normal behavior.

2.3 Optimum Detection

The detector monitors a process and, if it detects an anomaly, generates an alarm. It monitors the process through a sliding window containing w transitions. We refer to the content of the window as the monitored sample. The relative goodness by which the monitored sample is explained by the hypothesis that the monitored process is normal or intrusive determines whether an alarm is to be generated. This goodness of the explanation of a sample represented by the probability of that sample given the hypothesis. Figure 8 shows how the probability density function of the monitored sample depends on the hypothesis of normal behavior: the normalized mean of this function is the same as that of the characteristics, and the variance is greater by a value of $1/w^2$. The corresponding conditional density function for the hypothesis of intrusive behavior is shown in Fig. 9. The domain of these functions is the multi-dimensional space of monitored samples. If we represent this space by a line, the functions appear as in Fig. 10, where $|D|$, which is related to the covariance matrix of the normal density functions, is defined in the next section.

The conditional density functions of the monitored sample determine the optimum detection algorithm. As defined above, an optimum algorithm maximizes the true alarm rate under the constraint that the fraction of false alarms does not exceed a given value. It will be shown in the next section that the optimum algorithm can be defined by the locus of the points where the ratio of the two conditional functions is equal to a given factor. The monitored samples that fall on or outside this ellipsoid trigger an alarm, as indicated in Fig. 10.

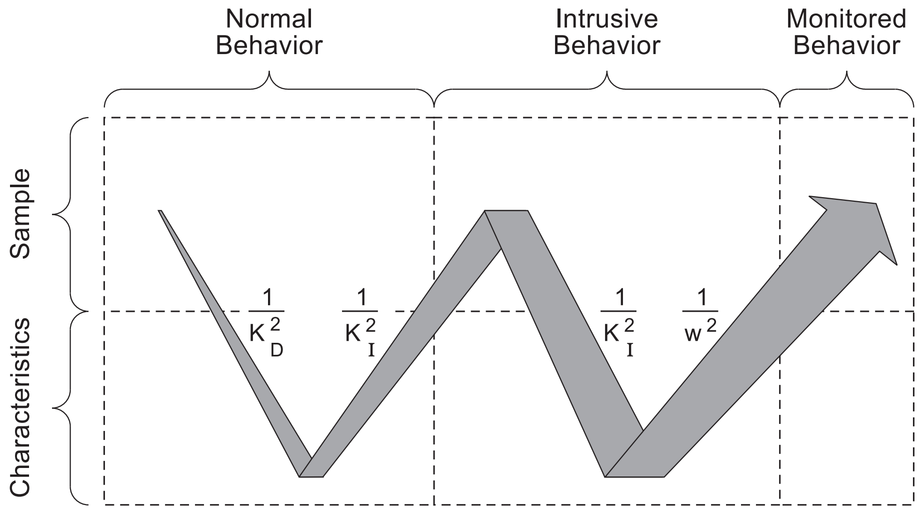

Figure 9: Uncertainty in the monitored sample of intrusive behavior.

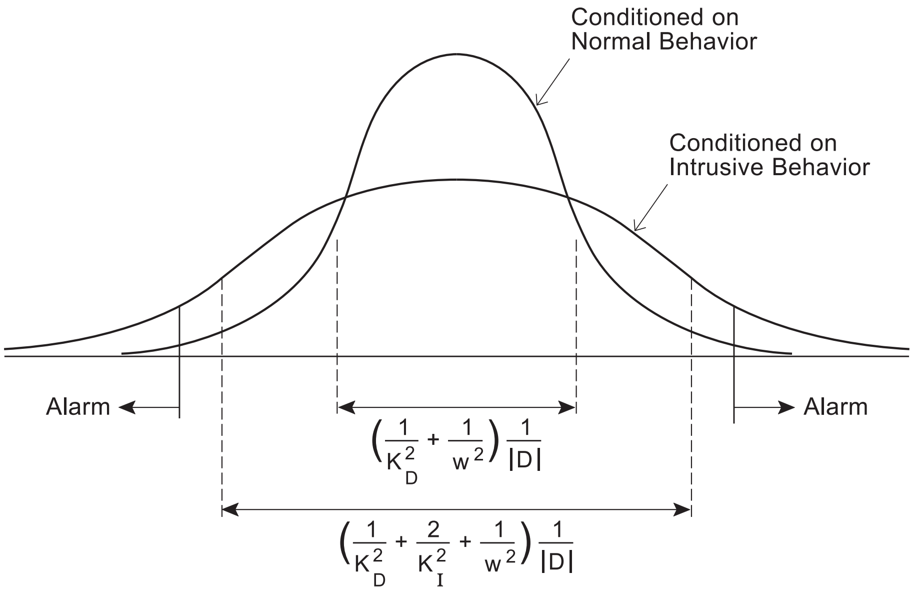

Figure 10: Conditional density functions of the monitored sample.

3 Analysis

In this section we derive the optimum detection algorithm from a sample of normal behavior and a given value of the masquerading factor, K_I. Without loss of generality we assume that there are no transient or null states in the sample of normal behavior, where transient and null are formally defined in [5]. As, based on this assumption, the mean time to the next occurrence of any state is finite, we further assume that the last state in the sample is the same as the first.

We now show how any sample in the form of a sequence of states can be used to derive a Markov model. Let K denote the number of state transitions in the sample, R the number of distinct transitions, n_r the number of occurrences of transition r in the sample, for $r =$

$1, 2, \ldots, K$ and $\underline{n}$ the vector $(n_1, n_2, \ldots, n_R)$.

In the previous section it was described how the state diagram of a Markov model can be extracted from a sample, as in the example of Fig. 3. The model will be complete when the state-transition probabilities are specified. Let S denote the number of states and, for $s = 1, 2, \ldots, S$, $O(s)$ the set of transitions with initial state s. For every transition r in $O(s)$, let p_r denote the probability that the state following s is that indicated by r. From the transition probabilities we can compute the steady-state probabilities of the states [5]. Let P_s denote the steady-state probability of state s, $s = 1, 2, \ldots, S$. For all $r \in O(s)$, we refer to $P_s\, p_r$ as the frequency of state r and denote it by θ_r. We observe that the transition and state probabilities are uniquely specified by the transition frequencies

$$P_s = \sum_{r \in O(s)} \theta_r \quad s = 1, 2, \ldots, S$$

and

$$p_r = \frac{\theta_r}{P_s} \quad r \in O(s) \quad ; \qquad s = 1, 2, \ldots, S \quad .$$

Let $\underline{\theta}$ denote the vector $(\theta_1, \theta_2, \ldots, \theta_R)$. Note that $\sum_{r=1}^{R} \theta_r$ is identical to 1. Then, based on Baye's theorem, the conditional density function of transition frequencies is

$$f(\underline{\theta} \mid \underline{n}) \propto \prod_{i=1}^{R} \theta_i^{n_i} \quad ,$$

where $\propto$ denotes proportionality within a normalization constant. Converting the product to a sum, we get

$$f(\underline{\theta} \mid \underline{n}_N) \propto \exp\left\{\sum_{i=1}^{R} n_i \ln \theta_i\right\} \quad .$$

The Taylor series expansion of $\ln \theta_i$ about $\hat{\theta}_i$ results in

$$f(\underline{\theta}_N \mid \underline{n}_N) \propto \exp\left\{\sum_{i=1}^{R} n_i \left[\ln \hat{\theta}_i + \frac{\theta_i - \hat{\theta}_i}{\hat{\theta}_i} - \frac{(\theta_i - \hat{\theta}_i)^2}{2\hat{\theta}_i^2} + \frac{(\theta_i - \hat{\theta}_i)^3}{3\hat{\theta}_i^3} - \ldots\right]\right\} \quad .$$

As $\ln \hat{\theta}_i$ is a constant, it can be dropped from the proportionality. As $\sum_{i=0}^{R} \theta_i$ is identical to 1, we choose $\hat{\theta}_i$ equal to n_i/K to make $\sum_{i=0}^{R} n_i(\theta_i - \hat{\theta}_i)/\hat{\theta}_i$ identical to 0. For small deviations, that is for small values of $|\theta_i - \hat{\theta}_i|/\hat{\theta}_i$, we can discard terms in the series beyond quadratic to get

$$f(\underline{\theta} \mid \underline{n}) \propto \exp\left\{-\frac{1}{2/K^2} \sum_{i=1}^{R} \left(\theta_i - \frac{n_i}{K}\right)^2 / n_i\right\} \quad .$$

Let

$$\Lambda \stackrel{\text{def}}{=} \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{n_R} \end{pmatrix} \quad .$$

Then

$$f(\underline{\theta} \mid \underline{n}) \propto \exp\left\{-\frac{1}{2/K_D^2}\left(\underline{\theta} - \frac{\underline{n}}{K}\right) \Lambda \left(\underline{\theta} - \frac{\underline{n}}{K}\right)^T\right\} \quad ,$$

which has the same form as the multi-dimensional normal distribution. Even though vector $\underline{\theta}_N$ has R elements, the space of this distribution has a lower dimensionality because some of these elements can be expressed in terms of the others using the state diagram. More specifically, for $s = 1, 2, \ldots, S$, let $I(s)$ be the set of transitions with the final state equal to s. The sum of frequencies of transitions ending in a state must be equal to that of transitions starting at a state, which results in the following S equations:

$$\sum_{r \in I(s)} \theta_r = \sum_{r \in O(s)} \theta_r \ .$$

Note that for the sample of normal behavior, owing to our assumption that its first and last states are equal, this set of equations is exact. For the intruder's sample of normal behavior as well as the monitored sample, when the first and last states are not equal, the number of transitions out of the initial (into the final) state exceeds the number of transitions into (out of) it by one. Nevertheless, for a sufficiently large sample, the set of equations is a good approximation.

As one of the equations in this set is redundant, we have $\sum_{r=0}^{R} r = 1$. Hence we have R frequencies and S equations. In the example in the next section we can express S of the frequencies in terms of the remaining $R - S$. Using the state diagram of the model, such as the one shown in Fig. 11, we can show that this is always the case. The bold transitions represent a minimal-spanning tree which connects all the states and thus contains $S - 1$ transitions. We can draw a cut through each of these transitions that does not cut the other transitions in the minimum-spanning tree. As the sum of the frequencies of the transitions through the cut must be zero, the bold transition with the cut can be expressed in terms of other transitions, which are not bold. Conversely, none of the non-bold transitions can be expressed exclusively in terms of other non-bold transitions because any cut that partitions the state diagram must cut at least one bold transition. The normalization equation, $\sum_{r=0}^{R} \theta_r = 1$, allows one more transition to be expressed in terms of others. Let $\underline{\tilde{\theta}}_N$ denote the subvector containing the $R - S$ independent transition frequencies, in terms of which $\underline{\theta}_N$ can be expressed:

$$\underline{\theta}^T = A^T \underline{\tilde{\theta}}^T \quad .$$

Let $\underline{\tilde{n}}$ denote the subvector of n corresponding to $\underline{\tilde{\theta}}$. Substituting this transformation in the

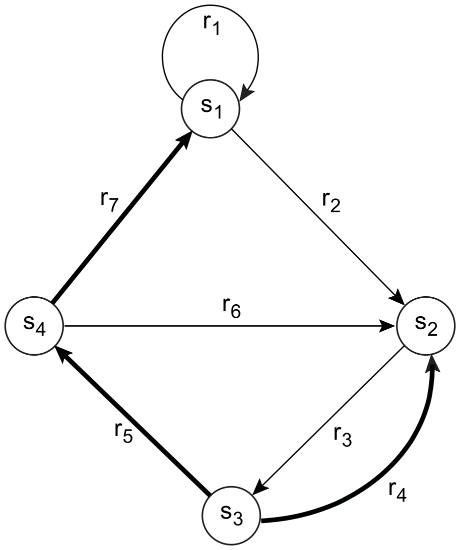

Figure 11: A minimum-spanning tree of the state diagram.

conditional density function, we get

$$f(\underline{\theta} \mid \underline{n}) \;\propto\; \exp\left\{-\frac{1}{2/K^2}\left(\underline{\tilde{\theta}}-\frac{\underline{\tilde{n}}}{K}\right) A\Lambda A^T \left(\underline{\tilde{\theta}}-\frac{\underline{\tilde{n}}}{K}\right)^T\right\} \quad .$$

Let g be the normal distribution for the transformation matrix A

$$g(\underline{z}, C) = \frac{1}{\sqrt{(2\pi)^{R-S}|C|}} \exp\left\{-\frac{\underline{\tilde{z}}C^{-1}\underline{\tilde{z}}^T}{2}\right\} \quad ,$$

where C is the covariance matrix [6]. We will then have

$$f(\underline{\theta} \mid \underline{n}) = g\left(\underline{\theta} - \frac{\underline{n}}{K}, K^{-2}D^{-1}\right) \quad ,$$

where D is defined as $A\Lambda A^T$.

We now derive the distribution of a sample of K states in terms of a model represented by a state-transition diagram and a parameter vector $\underline{\theta}$. As, within the constraint of the state-transition diagram the transitions are selected randomly, within a normalization constant the following proportionality holds:

$$f(\underline{n} \mid \underline{\theta}, K) \propto \prod_{r=0}^{R} \theta_r^{n_r} \quad ,$$

where the discrete distribution is approximated by a continuous distribution. As for $f(\underline{\theta} \mid \underline{n})$ we get

$$f(\underline{n} \mid \underline{\theta}, K) = g\left(\frac{\underline{n}}{K} - \underline{\theta}, w^{-2}D^{-1}\right) \quad .$$

Now we analyze the monitored sample of normal behavior in terms of the sample of normal behavior, based on Fig. 8. Let $\underline{n}_N$ and K_D denote, respectively, the sample of the normal behavior and its size, $\underline{\theta}_M$ the model parameters of the normal behavior, and $\underline{n}_M$ the monitored sample. As $\underline{n}_M$ depends on $\underline{n}_N$ only through $\underline{\theta}_N$, the normal random variables $\underline{\theta}_N - \underline{n}_N/K_D$ and $\underline{n}_M/w - \underline{\theta}_N$ are independent random variables and, because they are both normal random variables, as shown in [6], their sum is also a normal random variable with a mean equal to the sum of the means and a covariance matrix equal to the sum of covariance matrices:

$$f(\underline{n}_M \mid \underline{n}_N, w) = g\left(\frac{\underline{n}_M}{w} - \frac{\underline{n}_N}{K_D}, (K_D^{-2} + w^{-2})D^{-1}\right) \quad .$$

The probability density function for the monitored sample of intrusive behavior can be derived similarly; with reference to Fig. 9 we get

$$f(\underline{n}_M \mid \underline{n}_N, K_I, w) = g\left(\frac{\underline{n}_M}{w} - \frac{\underline{n}_N}{K_D}, (K_D^{-2} + 2K_I^{-2} + w^{-2})D^{-1}\right) \quad .$$

Now that we have obtained the conditional density functions of the monitored samples, we can develop the optimum detection algorithm. Our objective is to minimize the mean time to the detection of intrusive behavior under the constraint that the fraction of false alarms does not exceed a constant α, where $\alpha < 1$. The detection algorithm is defined by the values of $\underline{n}_M$ for which the alarm is triggered, which defines an alarm region. Clearly the more points we include in this region, the shorter the mean time to successful detection. So the optimum solution is to include all those points that satisfy the constraint.

To find the region that satisfies the constraint we find an expression for the value of the fraction of false alarms. Let $P(\mathrm{NB})$ and $P(\mathrm{IB})$ denote the *a priori* probabilities that the monitored process exhibits normal and intrusive behavior, respectively, and $P(\mathrm{NB} \mid \underline{\mathrm{n}}_{\mathrm{M}})$ and $P(\mathrm{IB} \mid \underline{\mathrm{n}}_{\mathrm{M}})$ the same probabilities given the monitored sample. Note that $P(\mathrm{IB})$ can be defined in real-time by the other detectors in the intrusion-detection system. According to Baye's theorem [7],

$$P(\mathrm{IB} \mid \underline{n}_M) = \frac{f(\underline{n}_M \mid \underline{n}_N, K_I, w)\, P(\mathrm{IB})}{f(\underline{n}_M \mid \underline{n}_N, K_I, w)\, P(\mathrm{IB}) \,+\, f(\underline{n}_M \mid \underline{n}_N, w)\, P(\mathrm{NB})}$$

and

$$P(\mathrm{NB} \mid \underline{n}_M) = \frac{f(\underline{n}_M \mid \underline{n}_N, w)\, P(\mathrm{NB})}{f(\underline{n}_M \mid \underline{n}_N, K_I, w)\, P(\mathrm{IB}) \,+\, f(\underline{n}_M \mid \underline{n}_N, w)\, P(\mathrm{N})} \quad .$$

The ratio of the above equations should satisfy the following inequality

$$\frac{P(\mathrm{NB} \mid \underline{n}_M)}{P(\mathrm{IB} \mid \underline{n}_M)} = \frac{f(\underline{n}_M \mid \underline{n}_N, w)\, P(\mathrm{NB})}{f(\underline{n}_M \mid \underline{n}_N, K_I, w)\, P(\mathrm{IB})} < \frac{\alpha}{1-\alpha} \quad ,$$

which defines the constraint of the optimization. Substituting for the conditional density functions, we get

$$b \exp\left\{ -\frac{a}{2} \left(\frac{\underline{\tilde{n}}_M}{w} - \frac{\underline{\tilde{n}}_N}{K_D} \right) D \left(\frac{\underline{\tilde{n}}_M}{w} - \frac{\underline{\tilde{n}}_N}{K_D} \right)^T \right\} < \frac{\alpha P(\mathrm{IB})}{(1-\alpha) P(\mathrm{NB})} \quad ,$$

where

$$a \stackrel{\text{def}}{=} \frac{1}{K_D^{-2} + w^{-2}} - \frac{1}{K_D^{-2} + 2K_I^{-2} + w^{-2}} > 0$$

and

$$b \stackrel{\text{def}}{=} \sqrt{\frac{K_D^{-2} + K_I^{-2} + w^{-2}}{K_D^{-2} + w^{-2}}} \quad .$$

Taking the log of both sides, we get the condition for activating an alarm:

$$\left(\frac{\underline{\tilde{n}}_M}{w} - \frac{\underline{\tilde{n}}_N}{K_D} \right) D \left(\frac{\underline{\tilde{n}}_M}{w} - \frac{\underline{\tilde{n}}_N}{K_D} \right)^T > \frac{2}{a} \ln \left[\frac{b(1-\alpha) P(\mathrm{NB})}{\alpha P(\mathrm{IB})} \right] \quad .$$

The amount of computation required for detection can be found as follows. Every new state transition that appears in the window of the monitored process either causes no change in $\underline{\tilde{n}}_M$ or the value of one of its elements increases by 1 and the value of another drops by 1. Let $\underline{z}$ and $\underline{z} + \Delta \underline{z}$ denote $(\underline{\tilde{n}}_M / w) - (\underline{\tilde{n}}_N / K_D)$ before and after the new transition, respectively. We have

$$(\underline{z} + \Delta \underline{z}) D (\underline{z} + \Delta \underline{z})^T = \underline{z} D \underline{z}^T + \Delta \underline{z} D (2\underline{z} + \Delta \underline{z})^T .$$

The first term has already been computed. The computation of $\Delta \underline{z}$ requires multiplying one row of D by $+1$ and another row by -1 and adding them, which is equivalent to $R - S$ subtractions. The computation of $2\underline{z} + \Delta \underline{z}$ requires $R - S$ 1-bit shifts and $R - S$ additions. The inner product of $\Delta \underline{z} D$ and $2\underline{z} + \Delta \underline{z}$ requires $R - S$ multiplications and $R - S - 1$ additions. The result is added to $\underline{z} D \underline{z}^T$ and compared with a constant. The total number of arithmetic operations is $5(R - S) + 1$.

4 Example

Here we work out the example given in Fig. 11. Based on the minimum-spanning tree selected, θ_4, θ_5, and θ_7 can be expressed as follows:

$$\begin{cases} \theta_4 &= \theta_3 - \theta_2 - \theta_6 \\ \theta_5 &= \theta_6 + \theta_2 \\ \theta_7 &= \theta_2 \end{cases} .$$

Now we use $\sum_{r=1}^{R} = 1$ to express θ_1 as follows:

$$\theta_1 = 1 - 2\theta_2 - 2\theta_3 - \theta_6 \quad .$$

The vector of independent variables, $\underline{\tilde{\theta}}$, is in this case $(\theta_2, \theta_3, \theta_6)$ and we have the following transformation:

$$\left(\underline{\theta} - \frac{\underline{n}}{K}\right)^T = A^T \left(\underline{\tilde{\theta}} - \frac{\underline{\tilde{n}}}{K}\right)^T \quad ,$$

where

$$A = \begin{pmatrix} -2 & 1 & 0 & -1 & 1 & 0 & 1 \\ -2 & 0 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 & 1 & 0 \end{pmatrix} \quad .$$

As an example, consider the sample $\underline{n}$ equal to $(20, 20, 40, 10, 30, 10, 20)$. Based on the equations in the previous section we get

$$D = \frac{1}{10} \begin{pmatrix} \frac{23}{6} & 1 & \frac{7}{3} \\ 1 & \frac{13}{4} & 0 \\ \frac{7}{3} & 0 & \frac{17}{6} \end{pmatrix} \quad .$$

5 Conclusion

The scheme presented here provides an optimum anomaly-detection algorithm based on a modeling approach that is accurate for a certain regime of process and detection parameters. The accuracy of the model depends on the validity of the assumption that, in all samples, the deviation in the number of occurrences of each state transition from its mean value is small. For a sample of normal behavior of reasonable length this assumption holds as long as none of the states are rare. It should also hold for the window of the transitions monitored by the detector. This implies that for a larger number of distinct transitions we need a larger detection window. A larger window, however, degrades the response time of the system. The scheme is therefore more attractive when the number of distinct transitions is not large and none of them are rare.

For an anomaly-detection algorithm to be efficient, it should not require a computational power more than a fraction of that required by the process being monitored. The computation per state transition required by the proposed algorithm is about $5(R-S)$ arithmetic operations, where S and R are the number of distinct states and transitions, respectively.

If, for a given process, the proposed algorithm meets the requirements of computational efficiency and small deviation, two advantages can be gained by applying it. First, it is an optimum algorithm in that it minimizes the mean time to detection of intrusive behavior under the constraint that the fraction of false alarms does not exceed a maximum value. Second, it allows the intrusive behavior to be characterized by a single scalar parameter, the length of the intruder's (hypothetical) sample of the normal behavior, which is referred to as the masquerading factor.

References

[1] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network Intrusion Detection," *IEEE Network*, May/June 1994, pp. 26–41.

[2] S. Forrest, S. A. Hofmeyr, and A. Somayaji, "Computer Immunology," *Communication of the ACM*, October 1997, pp. 88–96.

[3] M. Nassehi, "Characterizing Masqueraders for Intrusion Detection," submitted to *ESORICS 98*, Louvain-la-Neuve, Belgium, February 26, 1998.

[4] G. Owen, *Game Theory*, second ed. (Academic, New York, 1982).

[5] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operation Research Vol. I: Stochastic Processes and Operating Characteristics*, (McGraw-Hill, New York, 1982).

[6] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering* (Wiley, New York, 1965).

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[8] H. S. Teng, K. Chen and S. C-Y Lu, "Adaptive Real-time Anomaly Detection Using Inductively Generated Sequential Patterns," *Proc. IEEE Symp. on Res. in Sec. & Privacy*, pp. 278-84, Oakland, 1990.

[9] G. A. Stephen, *String Searching Algorithms* (World Scientific, Singapore, 1994).

[10] W. Feller, *An Introduction to Probability Theory and Its Applications*, Third Ed., (Wiley, New York, 1968).

[11] P. Helman and G. Liepins, "Statistical Foundations of Audit Trail Analysis for the Detection of Computer Misuse," *IEEE Trans. Software Eng.*, vol. 19, no. 9, Sept. 1993, pp. 886–901.

[12] A. V. Aho and M. J. Corasick, "Efficient String Matching: An Aid to Bibliographic Search," *Communications of the ACM*, vol. 18, no. 6, pp. 333-40, June 1975.