A Framework for the Evaluation of Intrusion Detection Systems

Alvaro A. Cárdenas John S. Baras Karl Seamon *
Department of Electrical and Computer Engineering
and The Institute of Systems Research
University of Maryland, College Park
{acardena,baras,kks}@isr.umd.edu

Abstract

Classification accuracy in intrusion detection systems (IDSs) deals with such fundamental problems as how to compare two or more IDSs, how to evaluate the performance of an IDS, and how to determine the best configuration of the IDS. In an effort to analyze and solve these related problems, evaluation metrics such as the Bayesian detection rate*, the* expected cost*, the* sensitivity *and the* intrusion detection capability *have been introduced. In this paper, we study the advantages and disadvantages of each of these performance metrics and analyze them in a unified framework. Additionally, we introduce the intrusion detection operating characteristic (IDOC) curves as a new IDS performance tradeoff which combines in an intuitive way the variables that are more relevant to the intrusion detection evaluation problem. We also introduce a formal framework for reasoning about the performance of an IDS and the proposed metrics against adaptive adversaries. We provide simulations and experimental results to illustrate the benefits of the proposed framework.*

1. Introduction

Consider a company that, in an effort to improve its information technology security infrastructure, wants to purchase either intrusion detector 1 ($I\mathcal{DS}_1$) or intrusion detector 2 ($I\mathcal{DS}_2$). Furthermore, suppose that the algorithms used by each IDS are kept private and therefore the only way to determine the performance of each IDS (unless some reverse engineering is done [15]) is through empirical tests determining how many intrusions are detected by each scheme while providing an acceptable level of false alarms. Suppose these tests show with high confidence that $I\mathcal{DS}_1$ detects one-tenth more attacks than $I\mathcal{DS}_2$ but at the cost of producing one hundred times more false alarms. The company needs to decide based on these estimates, which IDS will provide the best return of investment for their needs and their operational environment.

This general problem is more concisely stated as the intrusion detection evaluation problem, and its solution usually depends on several factors. The most basic of these factors are the *false alarm rate* and the *detection rate*, and their tradeoff can be intuitively analyzed with the help of the *receiver operating characteristic* (ROC) curve [16, 17, 35, 7, 14]. However, as pointed out in [3, 9, 10], the information provided by the detection rate and the false alarm rate alone might not be enough to provide a good evaluation of the performance of an IDS. Therefore, the evaluation metrics need to consider the environment the IDS is going to operate in, such as the maintenance costs and the hostility of the operating environment (the likelihood of an attack). In an effort to provide such an evaluation method, several performance metrics such as the *Bayesian detection rate* [3], *expected cost* [9], *sensitivity* [6] and *intrusion detection capability* [10], have been proposed in the literature.

Yet despite the fact that each of these performance metrics makes their own contribution to the analysis of intrusion detection systems, they are rarely applied in the literature when proposing a new IDS. It is our belief that the lack of widespread adoption of these metrics stems from two main reasons. Firstly, each metric is proposed in a different framework (e.g. information theory, decision theory, cryptography etc.) and in a seemingly ad hoc manner. Therefore an objective comparison between the metrics is very difficult.

The second reason is that the proposed metrics usually assume the knowledge of some uncertain parameters like the likelihood of an attack, or the costs of false alarms and missed detections. Moreover, these uncertain parameters can also change during the operation of an IDS. Therefore the evaluation of an IDS under some (wrongly) estimated parameters might not be of much value.

More importantly, there does not exist a security model

*This material is based upon work supported by the U.S. Army Research Office under Award No. DAAD19-01-1-0494 to the University of Maryland College Park.

Proceedings of the 2006 IEEE Symposium on Security and Privacy (S&P'06)
1081-6011/06 $20.00 © 2006 IEEE

for the evaluation of intrusion detection systems. Several researchers have pointed out the need to include the resistance against attacks as part of the evaluation of an IDS [25, 27, 11, 34, 29, 30, 13]. However, the traditional evaluation metrics are based on ideas mainly developed for non-security related fields and therefore, they do not take into account the role of an adversary and the evaluation of the system against this adversary. In particular, it is important to realize that when we borrow tools from other fields, they come with a set of assumptions that might not hold in an adversarial setting, because the first thing that the intruder will do is violate the sets of assumptions that the IDS is relying on for proper operation.

1.1. Our Contributions

In this paper, we introduce a framework for the evaluation of IDSs in order to address the concerns raised in the previous section. First, we identify the intrusion detection evaluation problem as a multi-criteria optimization problem. This framework will let us compare several of the previously proposed metrics in a unified manner. To this end, we recall that there are in general two ways to solve a multi-criteria optimization problem. The first approach is to combine the criteria to be optimized in a single optimization problem. We then show how the intrusion detection capability, the expected cost and the sensitivity metrics all fall into this category. The second approach to solve a multi-criteria optimization problem is to evaluate a tradeoff curve. We show how the Bayesian rates and the ROC curve analysis are examples of this approach.

To address the uncertainty of the parameters assumed in each of the metrics, we then present a graphical approach that allows the comparison of the IDS metrics for a wide range of uncertain parameters. For the single optimization problem we show how the concept of *isolines* can capture in a single value (the slope of the isoline) the uncertainties like the likelihood of an attack and the operational costs of the IDS. For the tradeoff curve approach, we introduce a new tradeoff curve we call the intrusion detector operating characteristic (IDOC). We believe the IDOC curve combines in a single graph all the relevant (and intuitive) parameters that affect the practical performance of an IDS.

Finally, we introduce a robust evaluation approach in order to deal with the adversarial environment the IDS is deployed in. In particular, we do not want to find the best performing IDS on average, but the IDS that performs best against the worst type of attacks. To that end we extend our graphical approach presented in section 4 to model the attacks against an IDS. In particular, we show how to find the best performing IDS against the worst type of attacks. This framework will allow us to reason about the security of the IDS evaluation and the proposed metric against adaptive adversaries.

In an effort to make this evaluation framework accessible to other researchers and in order to complement our presentation, we started the development of a software application available at [2] to implement the graphical approach for the expected cost and our new IDOC analysis curves. We hope this tool can grow to become a valuable resource for research in intrusion detection.

2. Notation and Definitions

In this section we present the basic notation and definitions which we use throughout this paper.

We assume that the input to an intrusion detection system is a feature-vector $\mathbf{x} \in \mathcal{X}$. The elements of $\mathbf{x}$ can include basic attributes like the duration of a connection, the protocol type, the service used etc. It can also include specific attributes selected with domain knowledge such as the number of failed logins, or if a superuser command was attempted. Examples of $\mathbf{x}$ used in intrusion detection are sequences of system calls [8], sequences of user commands [26], connection attempts to local hosts [12], proportion of accesses (in terms of TCP or UDP packets) to a given port of a machine over a fixed period of time [19] etc.

Let I denote whether a given instance $\mathbf{x}$ was generated by an intrusion (represented by $I = 1$ or simply I) or not (denoted as $I = 0$ or equivalently $\neg I$). Also let A denote whether the output of an IDS is an alarm (denoted by $A = 1$ or simply A) or not (denoted by $A = 0$, or equivalently $\neg A$). An IDS can then be defined as an algorithm $I\mathcal{DS}$ that receives a continuous data stream of computer event features $\mathbf{X} = \{\mathbf{x}[1], \mathbf{x}[2], \ldots, \}$ and classifies each input $\mathbf{x}[j]$ as being either a normal event or an attack i.e. $I\mathcal{DS} : \mathcal{X} \rightarrow \{A, \neg A\}$. In this paper we do not address how the IDS is designed. Our focus will be on how to evaluate the performance of a given IDS.

Intrusion detection systems are commonly classified as either *misuse* detection schemes or *anomaly* detection schemes. Misuse detection systems use a number of attack signatures describing attacks; if an event feature $\mathbf{x}$ matches one of the signatures, an alarm is raised. Anomaly detection schemes on the other hand rely on profiles or models of the normal operation of the system. Deviations from these established models raise alarms.

The empirical results of a test for an IDS are usually recorded in terms of how many attacks were detected and how many false alarms were produced by the IDS, in a data set containing both normal data and attack data. The percentage of alarms out of the total number of normal events monitored is referred to as the *false alarm rate* (or the *probability of false alarm*), whereas the percentage of detected attacks out of the total attacks is called the *detection rate* (or *probability of detection*) of the IDS. In general we de-

note the probability of false alarm and the probability of detection (respectively) as:

$$P_{FA} \equiv \Pr[A = 1|I = 0] \quad \text{and} \quad P_D \equiv \Pr[A = 1|I = 1] \quad (1)$$

These empirical results are sometimes shown with the help of the ROC curve; a graph whose x-axis is the false alarm rate and whose y-axis is the detection rate. The graphs of misuse detection schemes generally correspond to a single point denoting the performance of the detector. Anomaly detection schemes on the other hand, usually have a monitored statistic which is compared to a threshold τ in order to determine if an alarm should be raised or not. Therefore their ROC curve is obtained as a parametric plot of the probability of false alarm (P_{FA}) versus the probability of detection (P_D) (with parameter τ) as in [16, 17, 35, 7, 14].

3. Evaluation Metrics

In this section we first introduce metrics that have been proposed in previous work. Then we discuss how we can use these metrics to evaluate the IDS by using two general approaches, that is the expected cost and the tradeoff approach. In the expected cost approach, we give intuition of the expected cost metric by relating all the uncertain parameters (such as the probability of an attack) to a single line that allows the IDS operator to easily find the optimal tradeoff. In the second approach, we identify the main parameters that affect the quality of the performance of the IDS. This will allow us to later introduce a new evaluation method that we believe better captures the effect of these parameters than all previously proposed methods.

3.1. Background Work

Expected Cost. In this section we present the expected cost of an IDS by combining some of the ideas originally presented in [9] and [28]. The expected cost is used as an evaluation method for IDSs in order to assess the investment of an IDS in a given IT security infrastructure. In addition to the rates of detection and false alarm, the expected cost of an IDS can also depend on the hostility of the environment, the IDS operational costs, and the expected damage done by security breaches.

A quantitative measure of the consequences of the output of the IDS to a given event, which can be an intrusion or not are the costs shown in Table 1. Here $C(0,1)$ corresponds to the cost of responding as though there was an intrusion when there is none, $C(1,0)$ corresponds to the cost of failing to respond to an intrusion, $C(1,1)$ is the cost of acting upon an intrusion when it is detected (which can be defined as a negative value and therefore be considered as a profit for using the IDS), and $C(0,0)$ is the cost of not reacting to a non-intrusion (which can also be defined as a profit, or simply left as zero.)

State of the system	Detector's report	
	No Alarm (A=0)	Alarm (A=1)
No Intrusion ($I = 0$)	$C(0,0)$	$C(0,1)$
Intrusion ($A = 1$)	$C(1,0)$	$C(1,1)$

Table 1. Costs of the IDS reports given the state of the system

Adding costs to the different outcomes of the IDS is a way to generalize the usual tradeoff between the probability of false alarm and the probability of detection to a tradeoff between the *expected cost for a non-intrusion*

$$R(0,P_{FA}) \equiv C(0,0)(1-P_{FA}) + C(0,1)P_{FA}$$

and the *expected cost for an intrusion*

$$R(1,P_D) \equiv C(1,0)(1-P_D) + C(1,1)P_D$$

It is clear that if we only penalize errors of classification with unit costs (i.e. if $C(0,0) = C(1,1) = 0$ and $C(0,1) = C(1,0) = 1$) the expected cost for non-intrusion and the expected cost for intrusion become respectively, the false alarm rate and the detection rate.

The question of how to select the optimal tradeoff between the expected costs is still open. However, if we let the hostility of the environment be quantified by the *likelihood of an intrusion* $p \equiv \Pr[I = 1]$ (also known as the *base-rate* [3]), we can average the expected non-intrusion and intrusion costs to give the overall *expected cost of the IDS*:

$$\mathbf{E}[C(I,A)] = R(0,P_{FA})(1-p) + R(1,P_D)p \quad (2)$$

It should be pointed out that $R()$ and $\mathbf{E}[C(I,A)]$ are also known as the *risk* and *Bayesian risk* functions (respectively) in Bayesian decision theory.

Given an IDS, the costs from Table 1 and the likelihood of an attack p, the problem now is to find the optimal tradeoff between P_D and P_{FA} in such a way that $\mathbf{E}[C(I,A)]$ is minimized.

The Intrusion Detection Capability. The main motivation for introducing the *intrusion detection capability* C_{ID} as an evaluation metric originates from the fact that the costs in Table 1 are chosen in a subjective way [10]. Therefore the authors propose the use of the intrusion detection capability as an objective metric motivated by information theory:

$$C_{ID} = \frac{\mathbf{I}(I;A)}{\mathbf{H}(I)} \quad (3)$$

where $\mathbf{I}$ and $\mathbf{H}$ respectively denote the mutual information and the entropy [5]. The $\mathbf{H}(I)$ term in the denominator is

a normalizing factor so that the value of C_{ID} will always be in the $[0,1]$ interval. The intuition behind this metric is that by fine tuning an IDS based on C_{ID} we are finding the operating point that minimizes the uncertainty of whether an arbitrary input event $\mathbf{x}$ was generated by an intrusion or not.

The main drawback of C_{ID} is that it obscures the intuition that is to be expected when evaluating the performance of an IDS. This is because the notion of reducing the uncertainty of an attack is difficult to quantify in practical values of interest such as false alarms or detection rates. Information theory has been very useful in communications because the entropy and mutual information can be linked to practical quantities, like the number of bits saved by compression (source coding) or the number of bits of redundancy required for reliable communications (channel coding). However it is not clear how these metrics can be related to quantities of interest for the operator of an IDS.

The Base-Rate Fallacy and Predictive Value Metrics. In [3] Axelsson pointed out that one of the causes for the large amount of false alarms that intrusion detectors generate is the enormous difference between the amount of normal events compared to the small amount of intrusion events. Intuitively, the base-rate fallacy states that because the likelihood of an attack is very small, even if an IDS fires an alarm, the likelihood of having an intrusion remains relatively small. Formally, when we compute the posterior probability of intrusion (a quantity known as the *Bayesian detection rate*, or the *positive predictive value* (PPV)) given that the IDS fired an alarm, we obtain:

$$\begin{aligned}\text{PPV} &\equiv \Pr[I=1|A=1] \\ &= \frac{\Pr[A=1|I=1]\Pr[I=1]}{\Pr[A=1|I=1]\Pr[I=1]+\Pr[A=1|I=0]\Pr[I=0]} \\ &= \frac{P_D p}{(P_D-P_{FA})p+P_{FA}} \end{aligned} \tag{4}$$

Therefore, if the rate of incidence of an attack is very small, for example on average only 1 out of 10^5 events is an attack ($p=10^{-5}$), and if our detector has a probability of detection of one ($P_D=1$) and a false alarm rate of 0.01 ($P_{FA}=0.01$), then $\Pr[I=1|A=1]=0.000999$. That is on average, of 1000 alarms, only one would be a real intrusion.

It is easy to demonstrate that the PPV value is maximized when the false alarm rate of our detector goes to zero, even if the detection rate also tends to zero! Therefore as mentioned in [3] we require a trade-off between the PPV value and the *negative predictive value* (NPV):

$$\text{NPV} \equiv \Pr[I=0|A=0] = \frac{(1-p)(1-P_{FA})}{p(1-P_D)+(1-p)(1-P_{FA})} \tag{5}$$

3.2. Discussion

The concept of finding the optimal tradeoff of the metrics used to evaluate an IDS is an instance of the more general problem of multi-criteria optimization. In this setting, we want to maximize (or minimize) two quantities that are related by a tradeoff, which can be done via two approaches. The first approach is to find a suitable way of combining these two metrics in a single objective function (such as the expected cost) to optimize. The second approach is to directly compare the two metrics via a trade-off curve.

We therefore classify the above defined metrics into two general approaches that will be explored in the rest of this paper: the minimization of the expected cost and the tradeoff approach. We consider these two approaches as complimentary tools for the analysis of IDSs, each providing its own interpretation of the results.

Minimization of the Expected Cost. Let *ROC* denote the set of allowed (P_{FA},P_D) pairs for an IDS. The expected cost approach will include any evaluation metric that can be expressed as

$$r^* = \min_{(P_{FA},P_D)\in ROC} \mathbf{E}[C(I,A)] \tag{6}$$

where r^* is the expected cost of the IDS. Given $I\mathcal{DS}_1$ with expected cost r_1^* and an $I\mathcal{DS}_2$ with expected cost r_2^*, we can say $I\mathcal{DS}_1$ is better than $I\mathcal{DS}_2$ for our operational environment if $r_1^* < r_2^*$.

We now show how C_{ID}, and the tradeoff between the PPV and NPV values can be expressed as an expected costs problems. For the C_{ID} case note that the entropy of an intrusion $\mathbf{H}(I)$ is independent of our optimization parameters (P_{FA},P_D), therefore we have:

$$\begin{aligned}(P_{FA}^*,P_D^*) &= \arg\max_{(P_{FA},P_D)\in ROC} \frac{\mathbf{I}(I;A)}{\mathbf{H}(I)} \\ &= \arg\max_{(P_{FA},P_D)\in ROC} \mathbf{I}(I;A) \\ &= \arg\min_{(P_{FA},P_D)\in ROC} \mathbf{H}(I|A) \\ &= \arg\min_{(P_{FA},P_D)\in ROC} \mathbf{E}[-\log\Pr[I|A]]\end{aligned}$$

It is now clear that C_{ID} is an instance of the expected cost problem with costs given by $C(i,j)=-\log\Pr[I=i|A=j]$. By finding the costs of C_{ID} we are making the C_{ID} metric more intuitively appealing, since any optimal point that we find for the IDS will have an explanation in terms of cost functions (as opposed to the vague notion of diminishing the uncertainty of the intrusions).

Finally, in order to combine the PPV and the NPV in an average cost metric, recall that we want to maximize both

$\Pr[I=1|A=1]$ and $\Pr[I=0|A=0]$. Our average gain for each operating point of the IDS is therefore

$$\omega_1 \Pr[I=I|A=1]\Pr[A=1] + \omega_2 \Pr[I=0|A=0]\Pr[A=0]$$

where ω_1 (ω_2) is a weight representing a preference towards maximizing PPV (NPV). This equation is equivalent to the minimization of

$$-\omega_1 \Pr[I=I|A=1]\Pr[A=1] - \omega_2 \Pr[I=0|A=0]\Pr[A=0] \quad (7)$$

Comparing equation (7) with equation (2), we identify the costs as being $C(1,1) = -\omega_1$, $C(0,0) = -\omega_2$ and $C(0,1) = C(1,0) = 0$. Relating the predictive value metrics (PPV and NPV) with the expected cost problem will allow us to examine the effects of the base-rate fallacy on the expected cost of the IDS in future sections.

IDS classification tradeoffs. An alternate approach in evaluating intrusion detection systems is to directly compare the tradeoffs in the operation of the system by a tradeoff curve, such as ROC, or DET curves [21] (a reinterpretation of the ROC curve where the y-axis is $1-P_D$, as opposed to P_D). As mentioned in [3], another tradeoff to consider is between the PPV and the NPV values. However, we do not know of any tradeoff curves that combine these two values to aid the operator in choosing a given operating point.

We point out in section 4.2 that a tradeoff between P_{FA} and P_D (as in the ROC curves) as well as a tradeoff between PPV and NPV can be misleading for cases where p is very small, since very small changes in the P_{FA} and NPV values for our points of interest will have drastic performance effects on the P_D and the PPV values. Therefore, in the next section we introduce the IDOC as a new tradeoff curve between P_D and PPV.

4. Graphical Analysis

We now introduce a graphical framework that allows the comparison of different metrics in the analysis and evaluation of IDSs. This graphical framework can be used to adaptively change the parameters of the IDS based on its actual performance during operation. The framework also allows for the comparison of different IDSs under different operating environments.

Throughout this section we use one of the ROC curves analyzed in [9] and in [10]. Mainly the ROC curve describing the performance of the COLUMBIA team intrusion detector for the 1998 DARPA intrusion detection evaluation [18]. Unless otherwise stated, we assume for our analysis the base-rate present in the DARPA evaluation which was $p = 6.52 \times 10^{-5}$.

4.1. Visualizing the Expected Cost: The Minimization Approach

The biggest drawback of the expected cost approach is that the assumptions and information about the likelihood of attacks and costs might not be known a priori. Moreover, these parameters can change dynamically during the system operation. It is thus desirable to be able to tune the uncertain IDS parameters based on feedback from its actual system performance in order to minimize $\mathbf{E}[C(I,A)]$.

We select the use of ROC curves as the basic 2-D graph because they illustrate the behavior of a classifier without regard to the uncertain parameters, such as the base-rate p and the operational costs $C(i,j)$. Thus the ROC curve decouples the classification performance from these factors [24]. ROC curves are also general enough such that they can be used to study anomaly detection schemes and misuse detection schemes (a misuse detection scheme has only one point in the ROC space).

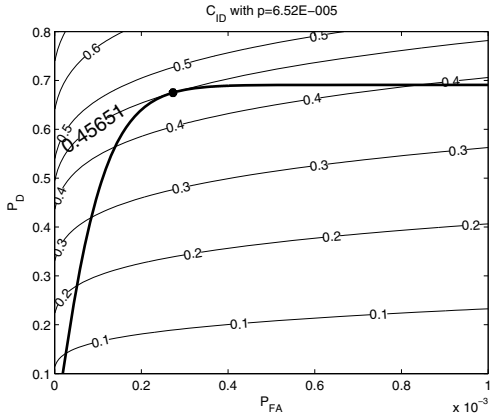


Figure 1. Isoline projections of C_{ID} onto the ROC curve. The optimal C_{ID} value is $C_{ID} = 0.4565$. The associated costs are $C(0,0) = 3 \times 10^{-5}$, $C(0,1) = 0.2156$, $C(1,0) = 15.5255$ and $C(1,1) = 2.8487$. The optimal operating point is $P_{FA} = 2.76 \times 10^{-4}$ and $P_D = 0.6749$.

In the graphical framework, the relation of these uncertain factors with the ROC curve of an IDS will be reflected in the *isolines* of each metric, where isolines refer to lines that connect pairs of false alarm and detection rates such that any point on the line has equal expected cost. The evaluation of an IDS is therefore reduced to finding the point of the ROC curve that intercepts the optimal isoline of the metric (for signature detectors the evaluation corresponds to finding the isoline that intercepts their single point in the ROC space and the point (0,0) or (1,1)). In Figure 1 we can

see as an example the isolines of C_{ID} intercepting the ROC curve of the 1998 DARPA intrusion detection evaluation.

One limitation of the C_{ID} metric is that it specifies the costs $C(i,j)$ a priori. However, in practice these costs are rarely known in advance and moreover the costs can change and be dynamically adapted based on the performance of the IDS. Furthermore the nonlinearity of C_{ID} makes it difficult to analyze the effect different p values will have on C_{ID} in a single 2-D graph. To make the graphical analysis of the cost metrics as intuitive as possible, we will assume from now on (as in [9]) that the costs are tunable parameters and yet once a selection of their values is made, they are constant values. This new assumption will let us at the same time see the effect of different values of p in the expected cost metric.

Under the assumption of constant costs, we can see that the isolines for the expected cost $\mathbf{E}[C(I,A)]$ are in fact straight lines whose slope depends on the ratio between the costs and the likelihood ratio of an attack. Formally, if we want the pair of points (P_{FA1},P_{D1}) and (P_{FA2},P_{D2}) to have the same expected cost, they must be related by the following equation [23, 31, 24]:

$$m_{C,p} \equiv \frac{P_{D2}-P_{D1}}{P_{FA1}-P_{FA2}} = \frac{1-p}{p}\frac{C(0,1)-C(0,0)}{C(1,0)-C(1,1)} = \frac{1-p}{p}\frac{1}{C} \tag{8}$$

where in the last equality we have implicitly defined C to be the ratio between the costs, and $m_{C,p}$ to be the slope of the isoline. The set of isolines of $\mathbf{E}[C(I,A)]$ can be represented by

$$ISO_E = \{m_{C,p} \times P_{FA} + b : b \in [0,1]\} \tag{9}$$

For fixed C and p, it is easy to prove that the optimal operating point of the ROC is the point where the ROC intercepts the isoline in ISO_E with the largest b (note however that there are ROC curves that can have more than one optimal point.) The optimal operating point in the ROC is therefore determined only by the slope of the isolines, which in turn is determined by p and C. Therefore we can readily check how changes in the costs and in the likelihood of an attack will impact the optimal operating point.

Effect of the Costs. In Figure 2, consider the operating point corresponding to $C = 58.82$, and assume that after some time, the operators of the IDS realize that the number of false alarms exceeds their response capabilities. In order to reduce the number of false alarms they can increase the cost of a false alarm $C(0,1)$ and obtain a second operating point at $C = 10$. If however the situation persists (i.e. the number of false alarms is still much more than what operators can efficiently respond to) and therefore they keep increasing the cost of a false alarm, there will be a *critical slope* m^c such that the intersection of the ROC and the isoline with slope m^c will be at the point $(P_{FA},P_D) = (0,0)$.

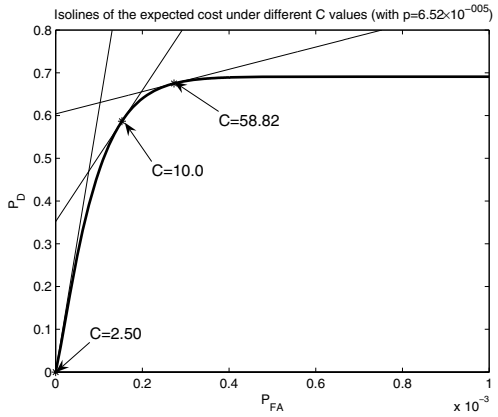


Figure 2. As the cost ratio C increases, the slope of the optimal isoline decreases

The interpretation of this result is that we should not use the IDS being evaluated since its performance is not good enough for the environment it has been deployed in. In order to solve this problem we need to either change the environment (e.g. hire more IDS operators) or change the IDS (e.g. shop for a more expensive IDS).

The Base-Rate Fallacy Implications on the Costs of an IDS. A similar scenario occurs when the likelihood of an attack changes. In Figure 3 we can see how as p decreases, the optimal operating point of the IDS tends again to $(P_{FA},P_D) = (0,0)$ (again the evaluator must decide not to use the IDS for its current operating environment). Therefore, for small base-rates the operation of an IDS will be cost efficient only if we have an appropriate large C^* such that $m_{C^*,p^*} \leq m^c$. A large C^* can be explained if cost of a false alarm much smaller than the cost of a missed detection: $C(1,0) \gg C(0,1)$ (e.g. the case of a government network that cannot afford undetected intrusions and has enough resources to sort through the false alarms).

Generalizations. This graphical method of cost analysis can also be applied to other metrics in order to get some insight into the expected cost of the IDS. For example in [6], the authors define an IDS with input space $\mathcal{X}$ to be $\sigma - sensitive$ if there exists an efficient algorithm with the same input space $\mathcal{E}: \mathcal{X} \rightarrow \{\neg A, A\}$, such that $P_D^{\mathcal{E}} - P_{FA}^{\mathcal{E}} \geq \sigma$. This metric can be used to find the optimal point of an ROC because it has a very intuitive explanation: as long as the rate of detected intrusions increases faster than the rate of false alarms, we keep moving the operating point of the IDS towards the right in the ROC. The optimal sensitivity problem for an IDS with a receiver operating characteristic *ROC*

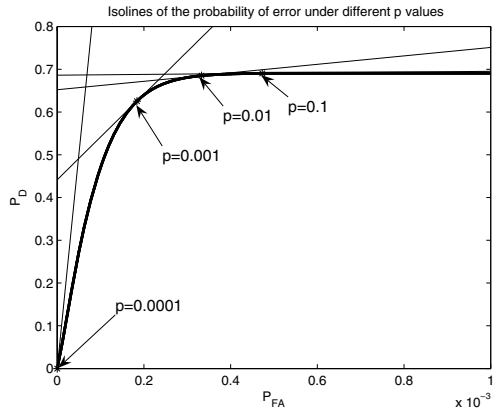


Figure 3. As the base-rate p decreases, the slope of the optimal isoline increases

is thus:

$$\max_{(P_{FA},P_D)\in ROC} P_D - P_{FA} \tag{10}$$

It is easy to show that this optimal sensitivity point is the same optimal point obtained with the isolines method for $m_{C,p} = 1$ (i.e. $C = (1-p)/p$).

4.2. The Intrusion Detector Operating Characteristic: The Tradeoff Approach

Although the graphical analysis introduced so far can be applied to analyze the cost efficiency of several metrics, the intuition for the tradeoff between the PPV and the NPV is still not clear. Therefore we now extend the graphical approach by introducing a new pair of isolines, those of the PPV and the NPV metrics.

Lemma 1 *Two sets of points (P_{FA1}, P_{D1}) and (P_{FA2}, P_{D2}) have the same PPV value if and only if*

$$\frac{P_{FA2}}{P_{D2}} = \frac{P_{FA1}}{P_{D1}} = \tan\theta \tag{11}$$

where θ is the angle between the line $P_{FA} = 0$ and the isoline. Moreover the PPV value of an isoline at angle θ is

$$PPV_{\theta,p} = \frac{p}{p + (1-p)\tan\theta} \tag{12}$$

Similarly, two set of points (P_{FA1}, P_{D1}) and $(P_{FA2}, PD2)$ have the same NPV value if and only if

$$\frac{1-P_{D1}}{1-P_{FA1}} = \frac{1-P_{D2}}{1-P_{FA2}} = \tan\phi \tag{13}$$

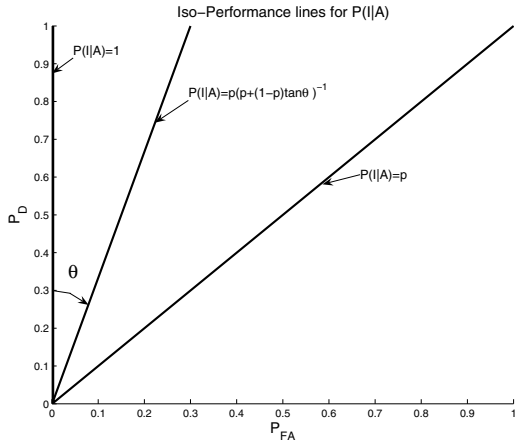


Figure 4. The PPV isolines in the ROC space are straight lines that depend only on θ. The PPV values of interest range from 1 to p

where ϕ is the angle between the line $P_D = 1$ and the isoline. Moreover the NPV value of an isoline at angle ϕ is

$$NPV_{\phi,p} = \frac{1-p}{p(\tan\phi - 1) + 1} \tag{14}$$

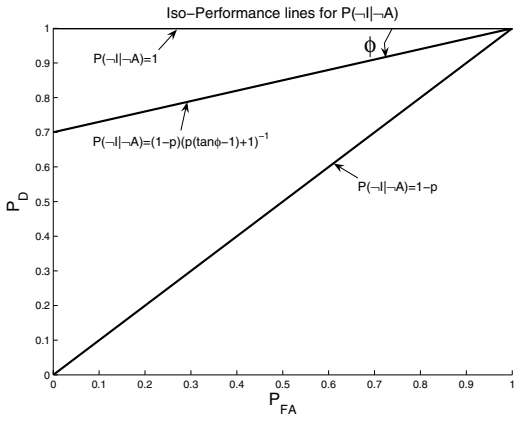


Figure 5. The NPV isolines in the ROC space are straight lines that depend only on ϕ. The NPV values of interest range from 1 to $1-p$

Figures 4 and 5 show the graphical interpretation of Lemma 1. It is important to note the range of the PPV and NPV values as a function of their angles. In particular notice that as θ goes from $0°$ to $45°$ (the range of interest), the

value of PPV changes from 1 to p. We can also see from figure 5 that as ϕ ranges from 0° to 45°, the NPV value changes from one to $1-p$. Therefore if p is very small $NPV \approx 1$. As shown in Figure 6, it turns out that the most relevant metrics to use for a tradeoff in the performance of an IDS are PPV and P_D.

However, even when you select as tradeoff parameters the PPV and P_D values, the isoline analysis shown in Figure 6 has still one deficiency when compared with the isoline performance analysis of the previous section, and it is the fact that there is no efficient way to represent how the PPV changes with different p values. In order to solve this problem we introduce the Intrusion Detector Operating Characteristic (IDOC) as a graph that shows how the two variables of interest: P_D and $\Pr[I = 1|A = 1]$ (the PPV value) are related under different base-rate values of interest. An example of an IDOC curve is presented in Figure (7). Although we show p varying substantially in this figure, the final choice for the uncertainty region of p is the choice of the user. Also note, for comparison purposes, that the IDOC curve of a classifier that performs random guessing is just a vertical line intercepting the x-axis at p, since $\Pr[I = 1|A = 1] = p$.

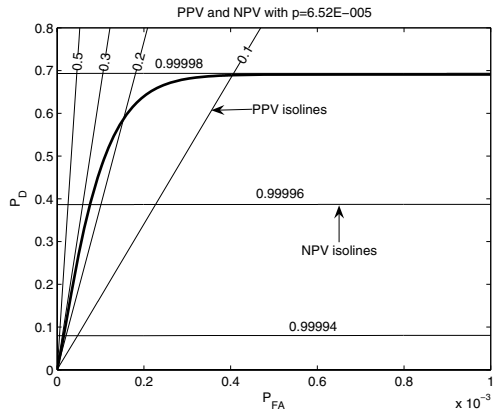


Figure 6. PPV and NPV isolines for the ROC of interest.

We believe that the IDOC provides a better way to evaluate IDS systems than most of the other previously proposed metrics. The ROC curve analysis alone does not take into account the estimated value of p. Furthermore, the operating points for an ROC might lead to misleading results as we do not know how to interpret intuitively very low false alarm rates where the precision might be misleading, e.g. is $P_{FA} = 10^{-3}$ much better than $P_{FA} = 5 \times 10^{-3}$? This same logic applies to the study of PPV vs NPV as we cannot interpret precisely small variations in NPV values, e.g. is $NPV = 0.9998$ much better than $NPV = 0.99975$? On the other hand in the IDOC curve we are comparing tradeoffs that are easier to interpret.

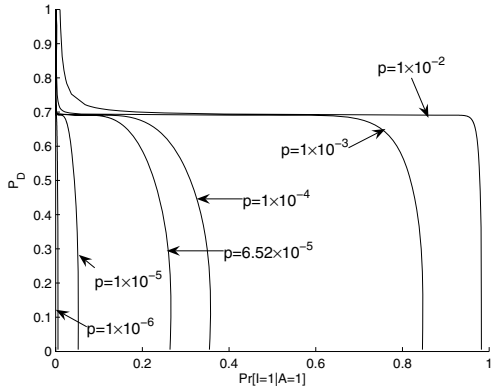


Figure 7. IDOC for the ROC of Figure 6.

5. Threat Models and Security Properties of the Evaluation

Traditional evaluation of intrusion detection schemes assumes that an intruder behaves similarly before and after the implementation of the IDS (i.e. a non-adaptive intruder). Now consider an intruder who adapts its attack when it faces a target system which hosts a given IDS.

For our evaluation analysis we were assuming three quantities that can be, up to a certain extent, controlled by the intruder. They are the base-rate p, the false alarm rate P_{FA} and the detection rate P_D. The base-rate can be modified by controlling the frequency of attacks. The perceived false alarm rate can be increased if the intruder finds a flaw in any of the signatures of an IDS that allows him to send maliciously crafted packets that trigger alarms at the IDS but that look benign to the IDS operator. Finally, the detection rate can be modified by the intruder with the creation of new attacks whose signatures do not match those of the IDS, or simply by evading the detection scheme, for example by the creation of a mimicry attack [34, 13].

In an effort towards understanding the advantage an intruder has by controlling these parameters, and to provide a robust evaluation framework, we now present a formal framework to reason about the robustness of an IDS evaluation method. Our work in this section is in some sense similar to the theoretical framework presented in [6], which was inspired by cryptographic models. However, we see two main differences in our work. First, we introduce the role of an adversary against the IDS, and thereby introduce a measure of robustness for the metrics. In the second place, our work is more practical and is applicable to more realistic

evaluation metrics. Furthermore we also provide examples of practical scenarios where our methods can be applied.

In order to be precise in our presentation, we need to extend the definitions introduced in section 2. For our modeling purposes we decompose the $I\mathcal{DS}$ algorithm into two parts: a *detector* $\mathcal{D}$ and a *decision maker* $\mathcal{DM}$. For the case of an anomaly detection scheme, $\mathcal{D}(\mathbf{x}[j])$ outputs the anomaly score $y[j]$ on input $\mathbf{x}[j]$ and $\mathcal{DM}$ represents the threshold that determines whether to consider the anomaly score as an intrusion or not, i.e. $\mathcal{DM}(y[j])$ outputs an alarm or it does not. For a misuse detection scheme, $\mathcal{DM}$ has to decide to use the signature to report alarms or decide that the performance of the signature is not good enough to justify its use and therefore will ignore all alarms (e.g. it is not cost-efficient to purchase the misuse scheme being evaluated).

Definition 1 An $I\mathcal{DS}$ algorithm is the composition of algorithms $\mathcal{D}$ (an algorithm from where we can obtain an ROC curve) and $\mathcal{DM}$ (an algorithm responsible for selecting an operating point). During operation, an $I\mathcal{DS}$ receives a continuous data stream of event features $\mathbf{x}[1], \mathbf{x}[2], \ldots$ and classifies each input $\mathbf{x}[j]$ by raising an alarm or not. Formally:[1]

$\underline{I\mathcal{DS}}(\mathbf{x})$
$y = \mathcal{D}(\mathbf{x})$
$A \leftarrow \mathcal{DM}(y)$
Output A (where $A \in \{0,1\}$)

$\diamondsuit$

We now study the performance of an IDS under an adversarial setting. We remark that our intruder model does not represent a single physical attacker against the IDS. Instead our model represents a collection of attackers whose average behavior will be studied under the worst possible circumstances for the IDS.

The first thing we consider, is the amount of information the intruder has. A basic assumption to make in an adversarial setting is to consider that the intruder knows everything that we know about the environment and can make inferences about the situation the same way as we can. Under this assumption we assume that the base-rate $\hat{p}$ estimated by the IDS, its estimated operating condition $(\hat{P}_{FA}, \hat{P}_D)$ selected during the evaluation, the original *ROC* curve (obtained from $\mathcal{D}$) and the cost function $C(I,A)$ are *public values* (i.e. they are known to the intruder).

We model the capability of an adaptive intruder by defining some confidence bounds. We assume an intruder can deviate $\hat{p} - \delta_l$, $\hat{p} + \delta_u$ from the expected $\hat{p}$ value. Also, based on our confidence in the detector algorithm and how hard we expect it to be for an intruder to evade the detector, or to create non-relevant false positives (this also models how the normal behavior of the system being monitored can produce new -previously unseen- false alarms), we define α and β as bounds to the amount of variation we can expect during the IDS operation from the false alarms and the detection rate (respectively) we expected, i.e. the amount of variation from $(\hat{P}_{FA}, \hat{P}_D)$ (although in practice estimating these bounds is not an easy task, testing approaches like the one described in [33] can help in their determination).

The intruder also has access to an oracle **Feature**$(\cdot,\cdot)$ that simulates an event to input into the IDS. **Feature**$(0,\zeta)$ outputs a feature vector modeling the normal behavior of the system that will raise an alarm with probability ζ (or a crafted malicious feature to only raise alarms in the case **Feature**$(0,1)$). And **Feature**$(1,\zeta)$ outputs the feature vector of an intrusion that will raise an alarm with probability ζ.

Definition 2 A $(\delta,\alpha,\beta) - intruder$ is an algorithm $\mathfrak{I}$ that can select its frequency of intrusions p_1 from the interval $\delta = [\hat{p} - \delta_l, \hat{p} + \delta_u]$. If it decides to attempt an intrusion, then with probability $p_2 \in [0,\beta]$, it creates an attack feature $\mathbf{x}$ that will go undetected by the IDS (otherwise this intrusion is detected with probability $\hat{P}_D$). If it decides not to attempt an intrusion, with probability $p_3 \in [0,\alpha]$ it creates a feature $\mathbf{x}$ that will raise a false alarm in the IDS

$\underline{\mathfrak{I}}(\delta,\alpha,\beta)$
Select $p_1 \in [\hat{p} - \delta_l, \hat{p} + \delta_u]$
Select $p_2 \in [0,\alpha]$
Select $p_3 \in [0,\beta]$
$I \leftarrow$ Bernoulli(p_1)
If $I = 1$
$B \leftarrow$ Bernoulli(p_3)
$\mathbf{x} \leftarrow$ **Feature**$(1, (min\{(1-B), \hat{P}_D\}))$
Else
$B \leftarrow$ Bernoulli(p_2)
$\mathbf{x} \leftarrow$ **Feature**$(0, max\{B, \hat{P}_{FA}\})$
Output (I,$\mathbf{x}$)

where Bernoulli(ζ) outputs a Bernoulli random variable with probability of success ζ.

Furthermore, if $\delta_l = p$ and $\delta_u = 1 - p$ we say that $\mathfrak{I}$ has the ability to make a *chosen-intrusion rate attack*.

$\diamondsuit$

We now formalize what it means for an evaluation scheme to be robust. We stress the fact that we are not analyzing the security of an IDS, but rather the security of the *evaluation* of an IDS, i.e. how confident we are that the IDS will behave during operation similarly to what we assumed in the evaluation.

[1] The usual arrow notation: $a \leftarrow \mathcal{DM}(y)$ implies that $\mathcal{DM}$ can be a probabilistic algorithm.

5.1. Robust Expected Cost Evaluation

We start with the general decision theoretic framework of evaluating the expected cost (per input) $\mathbf{E}[C(I,A)]$ for an IDS.

Definition 3 An evaluation method that claims the expected cost of an $I\mathcal{DS}$ is at most r is **robust** against a $(\delta,\alpha,\beta)-intruder$ if the expected cost of $I\mathcal{DS}$ during the attack ($\mathbf{E}^{\delta,\alpha,\beta}[C[I,A)]$) is no larger than r, i.e.

$$\mathbf{E}^{\delta,\alpha,\beta}[C[I,A)] = \sum_{i,a} C(i,a)\times$$
$$\Pr[\ (I,\mathbf{x}) \leftarrow \Im(\delta,\alpha,\beta);\ A \leftarrow I\mathcal{DS}(\mathbf{x}):\ I=i, A=a\ ] \le r$$

$\diamondsuit$

Now recall that the traditional evaluation framework finds an evaluation value r^* by using equation (6). So by finding r^* we are basically finding the best performance of an IDS and claiming the IDS is better than others if r^* is smaller than the evaluation of the other IDSs. In this section we claim that an IDS is better than others if its expected value under the worst performance is smaller than the expected value under the worst performance of other IDSs. In short

Traditional Evaluation Given a set of IDSs $\{I\mathcal{DS}_1, I\mathcal{DS}_2, \ldots, I\mathcal{DS}_n\}$ find the best expected cost for each:

$$r_i^* = \min_{(P_{FA}^{\alpha},P_D^{\beta})\in ROC_i} \mathbf{E}[C(I,A)] \tag{15}$$

Declare that the best IDS is the one with smallest expected cost r_i^*.

Robust Evaluation Given a set of IDSs $\{I\mathcal{DS}_1, I\mathcal{DS}_2, \ldots, I\mathcal{DS}_n\}$ find the best expected cost for each $I\mathcal{DS}_i$ when being under the attack of a $(\delta,\alpha_i,\beta_i)-intruder$[2]. Therefore we find the best IDS as follows:

$$r_i^{robust} = \min_{(P_{FA}^{\alpha_i},P_D^{\beta_i})\in ROC_i^{\alpha_i,\beta_i}} \max_{\Im(\delta,\alpha_i,\beta_i)} \mathbf{E}^{\delta,\alpha_i,\beta_i}[C(I,A)] \tag{16}$$

Several important questions can be raised by the above framework. In particular we are interested in finding the least upper bound r such that we can claim the evaluation of $I\mathcal{DS}$ to be *robust*. Another important question is how can we design an evaluation of $I\mathcal{DS}$ satisfying this least upper bound? Solutions to these questions are partially based on game theory.

[2]Note that different IDSs might have different α and β values. For example if $I\mathcal{DS}_1$ is an anomaly detection scheme then we can expect that the probability that new normal events will generate alarms α_1 is larger than the same probability α_2 for a misuse detection scheme $I\mathcal{DS}_2$.

Lemma 2 *Given an initial estimate of the base-rate $\hat{p}$, an initial ROC curve obtained from $\mathcal{D}$, and constant costs $C(I,A)$, the least upper bound r such that the expected cost evaluation of $I\mathcal{DS}$ is r-robust is given by*

$$r = R(0,\hat{P}_{FA}^{\alpha})(1-\hat{p}^{\delta}) + R(1,\hat{P}_D^{\beta})\hat{p}^{\delta} \tag{17}$$

where

$$R(0,\hat{P}_{FA}^{\alpha}) \equiv [C(0,0)(1-\hat{P}_{FA}^{\alpha}) + C(0,1)\hat{P}_{FA}^{\alpha}] \tag{18}$$

is the expected cost of $I\mathcal{DS}$ under no intrusion and

$$R(1,\hat{P}_D^{\beta}) \equiv [C(1,0)(1-\hat{P}_D^{\beta}) + C(1,1)\hat{P}_D^{\beta}] \tag{19}$$

is the expected cost of $I\mathcal{DS}$ under an intrusion, and $\hat{p}^{\delta}$, $\hat{P}_{FA}^{\alpha}$ and $\hat{P}_D^{\beta}$ are the solution to a zero-sum game between the intruder (the maximizer) and the IDS (the minimizer), whose solution can be found in the following way:

1. *Let (P_{FA},P_D) denote any points of the initial ROC obtained from $\mathcal{D}$ and let $ROC^{(\alpha,\beta)}$ be the ROC curve defined by the points $(P_{FA}^{\alpha},P_D^{\beta})$, where $P_D^{\beta} = P_D(1-\beta)$ and $P_{FA}^{\alpha} = \alpha + P_{FA}(1-\alpha)$.*

2. *Using $\hat{p}+\delta_u$ in the isoline method, find the optimal operating point (x_u,y_u)in $ROC^{(\alpha,\beta)}$ and using $\hat{p}-\delta_l$ in the isoline method, find the optimal operating point (x_l,y_l) in $ROC^{(\alpha,\beta)}$.*

3. *Find the points (x^*,y^*) in $ROC^{(\alpha,\beta)}$ that intersect the line*

$$y = \frac{C(1,0)-C(0,0)}{C(1,0)-C(1,1)} + x\frac{C(0,0)-C(0,1)}{C(1,0)-C(1,1)}$$

(under the natural assumptions $C(1,0) > R(0,x^*) > C(0,0)$, $C(0,1) > C(0,0)$ and $C(1,0) > C(1,1)$). If there are no points that intersect this line, then set $x^* = y^* = 1$.

4. *If $x^* \in [x_l,x_u]$ then find the base-rate parameter p^* such that the optimal isoline of Equation (9) intercepts $ROC^{(\alpha,\beta)}$ at (x^*,y^*) and set $\hat{p}^{\delta} = p^*$, $\hat{P}_{FA}^{\alpha} = x^*$ and $\hat{P}_D^{\beta} = y^*$.*

5. *Else if $R(0,x_u) < R(1,y_u)$ find the base-rate parameter p_u such that the optimal isoline of Equation (9) intercepts $ROC^{(\alpha,\beta)}$ at (x_u,y_u) and then set $\hat{p}^{\delta} = p_u$, $\hat{P}_{FA}^{\alpha} = x_u$ and $\hat{P}_D^{\beta} = y_u$. Otherwise, find the base-rate parameter p_l such that the optimal isoline of Equation (9) intercepts $ROC^{(\alpha,\beta)}$ at (x_l,y_l) and then set $\hat{p}^{\delta} = p_l$, $\hat{P}_{FA}^{\alpha} = x_l$ and $\hat{P}_D^{\beta} = y_l$.*

The proof of this lemma is very straightforward. The basic idea is that if the uncertainty range of p is large enough, the Nash equilibrium of the game is obtained by selecting the point intercepting equation (3). Otherwise one of the strategies for the intruder is always a dominant strategy of the game and therefore we only need to find which one is it: either $\hat{p}+\delta_u$ or $\hat{p}-\delta_l$. For most practical cases it will be $\hat{p}+\delta_u$. Also note that the optimal operating point in the original *ROC* can be found by obtaining $(\hat{P}_{FA},\hat{P}_D)$ from $(\hat{P}_{FA}^{\alpha},\hat{P}_D^{\beta})$.

5.2. Robust IDOC Evaluation

Similarly we can now also analyze the robustness of the evaluation done with the IDOC curves. In this case it is also easy to see that the worst attacker for the evaluation is an intruder $\mathfrak{I}$ that selects $p_1=\hat{p}-\delta_l$, $p_2=\alpha$ and $p_3=\beta$.

Corollary 3 *For any point $(P\hat{P}V,\hat{P}_D)$ corresponding to $\hat{p}$ in the IDOC curve, a $(\delta,\alpha,\beta)-$intruder can decrease the detection rate and the positive predictive value to the pair $(P\hat{P}V^{\delta,\alpha,\beta},\hat{P}_D^{\beta})$, where $\hat{P}^{\beta}=\hat{P}_D(1-\beta)$ and where*

$$P\hat{P}V^{\delta,\alpha,\beta}=\frac{P_D^{\beta}p-P^{\beta}\delta}{P_D^{\beta}p+P_{FA}^{\alpha}(1-p)+\delta P_{FA}^{\alpha}-\delta P_D^{\beta}} \qquad (20)$$

5.3. Example: Minimizing the Cost of a Chosen Intrusion Rate Attack

We now present an example that shows the generality of lemma 2 and also presents a compelling scenario of when does a probabilistic IDSs make sense. Assume an ad hoc network scenario similar to [20, 36, 32, 4] where nodes monitor and distribute reputation values of other nodes' behavior at the routing layer. The monitoring nodes report selfish actions (e.g. nodes that agree to forward packets in order to be accepted in the network, but then fail to do so) or attacks (e.g. nodes that modify routing information before forwarding it).

Now suppose that there is a network operator considering implementing a watchdog monitoring scheme to check the compliance of nodes forwarding packets as in [20]. The operator then plans an evaluation period of the method where trusted nodes will be the watchdogs reporting the misbehavior of other nodes. Since the detection of misbehaving nodes is not perfect, during the evaluation period the network operator is going to measure the consistency of reports given by several watchdogs and decide if the watchdog system is worth keeping or not.

During this trial period, it is of interest to selfish nodes to behave as deceiving as they can so that the neighboring watchdogs have largely different results and the system is not permanently established. As stated in [20] the watchdogs might not detect a misbehaving node in the presence of 1) ambiguous collisions, 2) receiver collisions, 3) limited transmission power, 4) false misbehavior, 5) collusion or 6) partial dropping. False alarms are also possible in several cases, for example when a node moves out of the previous node's listening range before forwarding on a packet. Also if a collision occurs while the watchdog is waiting for the next node to forward a packet, it may never overhear the packet being transmitted.

In this case, the detector algorithm $\mathcal{D}$ is the watchdog mechanism that monitors the medium to see if the packet was forwarded F or if it did not hear the packet being forwarded (unheard U) during a specified amount of time. Following [20] (where it is shown that the number of false alarms can be quite high) we assume that a given watchdog $\mathcal{D}$ has a false alarm rate of $\hat{P}_{FA}=0.5$ and a detection rate of $\hat{P}_D=0.75$. Given this detector algorithm, a (non-randomized) decision maker $\mathcal{DM}$ has to be one of the following rules (where intuitively, h_3 is the more appealing):

$$\begin{array}{ll} h_1(F)=0 & h_1(U)=0 \\ h_2(F)=1 & h_2(U)=0 \\ h_3(F)=0 & h_3(U)=1 \\ h_4(F)=1 & h_4(U)=1 \end{array}$$

Now notice that since the operator wants to check the consistency of the reports, the selfish nodes will try to maximize the probability of error (i.e. $C(0,0)=C(1,1)=0$ and $C(0,1)=C(1,0)=1$) of any watchdog with a chosen intrusion rate attack. As stated in lemma 2, this is a zero-sum game where the adversary is the maximizer and the watchdog is the minimizer. The matrix of this game is given in Table 2.

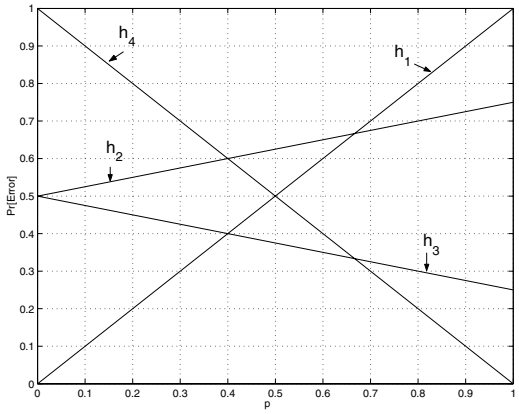


Figure 8. Probability of error for h_i vs. p

It is a well known fact that in order to achieve a Nash equilibrium of the game, the players should consider mixed strategies (i.e. consider probabilistic choices). For our

	h_0	h_1	h_2	h_3
$I=0$	$R(0,0)$	$R(0,\hat{P}_D)$	$R(0,\hat{P}_{FA})$	$R(0,1)$
$I=1$	$R(1,0)$	$R(1,\hat{P}_{FA})$	$R(1,\hat{P}_D)$	$R(1,1)$

Table 2. Matrix for the zero-sum game theoretic formulation of the detection problem

example the optimal mixed strategy for the selfish node (see Figure 8) is to drop a packet with probability $p^* = \hat{P}_{FA}/(\hat{P}_{FA1}+\hat{P}_D)$. On the other hand the optimal strategy for $\mathcal{DM}$ is to select h_3 with probability $1/(\hat{P}_{FA}+\hat{P}_D)$ and h_1 with probability $(\hat{P}_{FA}-(1-\hat{P}_D))/(\hat{P}_{FA}-(1-\hat{P}_D)+1)$. This example shows that sometimes in order to minimize the probability of error (or any general cost) against an adaptive attacker, $\mathcal{DM}$ has to be a probabilistic algorithm.

Lemma 2 also presents a way to get this optimal point from the ROC, however it is not obvious at the beginning how to get the same results, as there appear to be only three points in the ROC: $(P_{FA}=0,P_D=0)$ (by selecting h_1), $(\hat{P}_{FA}=1/2,\hat{P}_D=3/4)$ (by selecting h_3) and $(P_{FA}=1,P_D=1)$ (by selecting h_4). The key property of ROC curves to remember is that the (optimal) ROC curve is a continuous and concave function [23], and that in fact, the points that do not correspond to deterministic decisions are joined by a straight line whose points can be achieved by a mixture of probabilities of the extreme points. In our case, the line $y=1-x$ intercepts the (optimal) ROC at the optimal operating points $\hat{P}^*_{FA}=\hat{P}_{FA}/(\hat{P}_D+\hat{P}_{FA})$ and $\hat{P}^*_D=\hat{P}_D/(\hat{P}_{FA}+\hat{P}_D)$ (see Figure 9). Also note that p^* is the value required to make the slope of the isoline parallel to the ROC line intersecting (P^*_{FA},P^*_D).

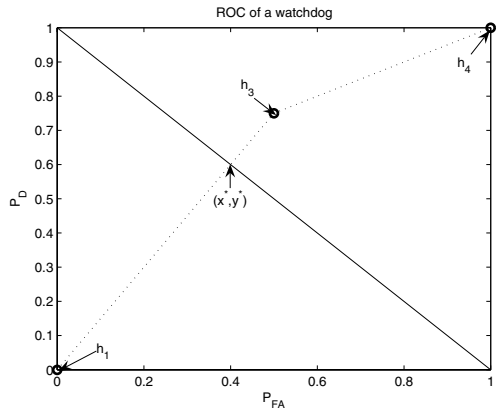


Figure 9. The optimal operating point

The optimal strategy for the intruder is therefore $p^*=2/5$, while the optimal strategy for $\mathcal{DM}$ is to select h_1 with probability $1/5$ and h_3 with probability $4/5$. In the robust operating point we have $P^*_{FA}=2/5$ and $P^*_D=3/5$. Therefore, after fixing $\mathcal{DM}$, it does not matter if p deviates from p^* because we are guaranteed that the probability of error will be no worse (but no better either) than $2/5$, therefore the IDS can be claimed to be $2/5$-robust.

5.4. Example: Robust Evaluation of IDSs

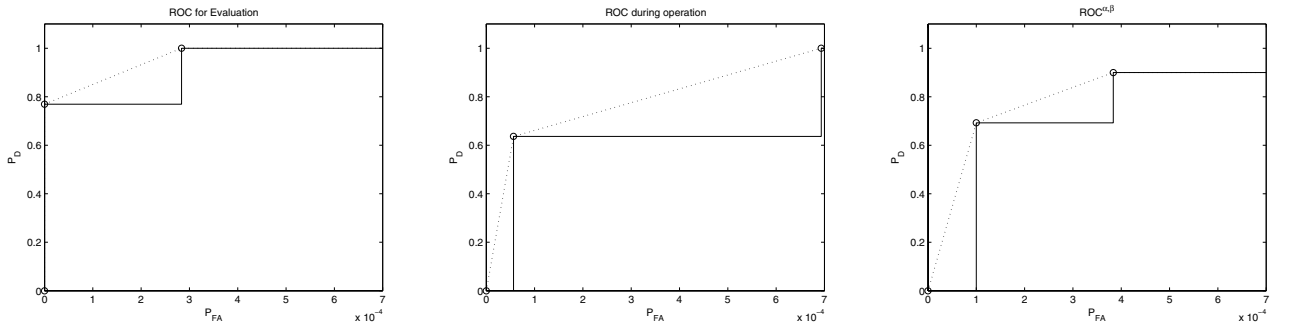
As a second example, we chose to perform an intrusion detection experiment with the 1998 MIT/Lincoln Labs data set [1]. Although several aspects of this data set have been criticized in [22], we still chose it for two main reasons. On one hand, it has been (and arguably still remains) the most used large-scale data set to evaluate IDSs. In the second place we are not claiming to have a better IDS to detect attacks and then proving our claim with its good performance in the MIT data set (a feat that would require further testing in order to be assured on the quality of the IDS). Our aim on the other hand is to illustrate our methodology, and since this data set is publicly available and has been widely studied and experimented with (researchers can in principle reproduce any result shown in a paper), we believe it provides the basic background and setting to exemplify our approach.

Of interest are the Solaris system log files, known as BSM logs. The first step of the experiment was to record every instance of a program being executed in the data set. Next, we created a very simple tool to perform buffer overflow detection. To this end, we compared the buffer length of each execution with a buffer threshold, if the buffer size of the execution was larger than the threshold we report an alarm.

We divided the data set into two sets. In the first one (weeks 6 and 7), our IDS performs very well and thus we assume that this is the "evaluation" period. The previous three weeks were used as the period of operation of the IDS. Figure 10(a)[3] shows the results for the "evaluation" period when the buffer threshold ranges between 64 and 780. The dotted lines represent the suboptimal points of the ROC or equivalently the optimal points that can be achieved through randomization. For example the dotted line of Figure 10(a) can be achieved by selecting with probability λ the detector with threshold 399 and with probability $1-\lambda$ the detector with threshold 773 and letting λ range from zero to one.

During the evaluation weeks there were 81108 executions monitored and 13 attacks, therefore $\hat{p}=1.6\times10^{-4}$.

[3] Care must always be taken when looking at the results of ROC curves due to the "unit of analysis" problem [22]. For example comparing the ROC of Figure 10(a) with the ROC of [14] one might arrive to the erroneous conclusion that the buffer threshold mechanism produces an IDS that is better than the more sophisticated IDS based on Bayesian networks. The difference lies in the fact that we are monitoring the execution of *every program* while the experiments in [14] only monitor the attacked programs (`eject`, `fbconfig`, `fdformat` and `ps`). Therefore although we raise more false alarms, our false alarm rate (number of false alarms divided by the total number of honest executions) is smaller.



(a) Original ROC obtained during the evaluation period

(b) Effective ROC during operation time

(c) Original ROC under adversarial attack $ROC^{\alpha,\beta}$

Figure 10. Robust expected cost evaluation

Assuming that our costs (per execution) are $C(0,0) = C(1,1) = 0$, $C(1,0) = 850$ and $C(0,1) = 100$ we find that the slope given by equation 8 is $m_{C,\hat{p}} = 735.2$, and therefore the optimal point is $(2.83 \times 10^{-4}, 1)$, which corresponds to a threshold of 399 (i.e. all executions with buffer sizes bigger than 399 raise alarms). Finally, with these operating conditions we find out that the expected cost (per execution) of the IDS is $\mathbf{E}[C(I,A)] = 2.83 \times 10^{-2}$.

In the previous three weeks used as the "operation" period our buffer threshold does not perform as well, as can be seen from its ROC (shown in Figure 10(b).) Therefore if we use the point recommended in the evaluation (i.e. the threshold of 399) we get an expected cost of $\mathbf{E}^{\text{operation}}[C(I,A)] = 6.934 \times 10^{-2}$. Notice how larger the expected cost per execution is from the one we had evaluated. This is very noticeable in particular because the base-rate is smaller during the operation period ($\hat{p}^{\text{operation}} = 7 \times 10^{-5}$) and a smaller base-rate should have given us a smaller cost.

To understand the new ROC let us take a closer look at the performance of one of the thresholds. For example, the buffer length of 773 which was able to detect 10 out of the 13 attacks at no false alarm in Figure 10(a) does not perform well in Figure 10(b) because some programs such as `grep`, `awk`, `find` and `ld` were executed under normal operation with long string lengths. Furthermore, a larger percent of attacks was able to get past this threshold. This is in general the behavior modeled by the parameters α and β that the adversary has access to in our framework.

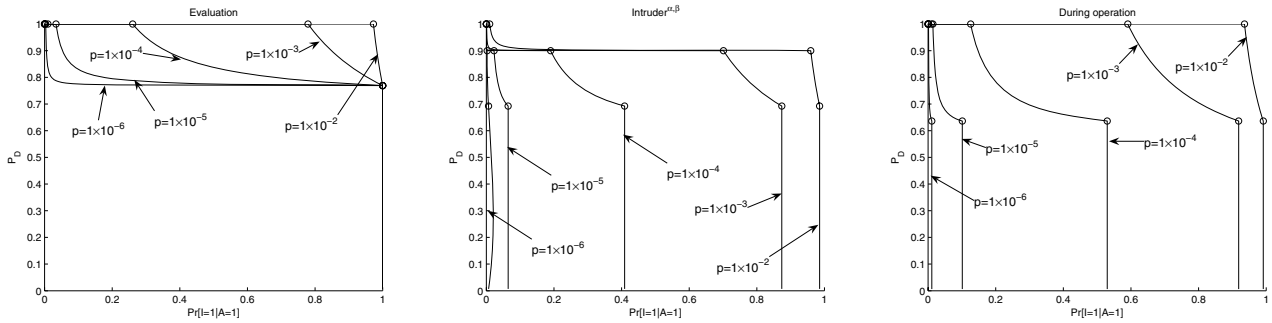
Let us begin the evaluation process from the scratch by assuming a $([1 \times 10^{-5}, 0], 1 \times 10^{-4}, 0.1) - intruder$, where $\delta = [1 \times 10^{-5}, 0]$ means the IDS evaluator believes that the base-rate during operation will be at most $\hat{p}$ and at least $\hat{p} - 1 \times 10^{-5}$. $\alpha = 1 \times 10^{-5}$ means that the IDS evaluator believes that new normal behavior will have the chance of firing an alarm with probability 1×10^{-5}. And $\beta = 0.1$ means that the IDS operator has estimated that ten percent of the attacks during operation will go undetected. With these parameters we get the $ROC^{\alpha,\beta}$ shown in Figure 10(c).

Note that in this case, p is bounded in such a way that the equilibrium of the game is achieved via a pure strategy. In fact, the optimal strategy of the intruder is to attack with frequency $\hat{p} + \delta_u$ (and of course, generate missed detections with probability β and false alarms with probability α) whereas the optimal strategy of $\mathcal{DM}$ is to find the point in $ROC^{\alpha,\beta}$ that minimizes the expected cost by assuming that the base-rate is $\hat{p} + \delta_u$.

The optimal point for the $ROC^{\alpha,\beta}$ curve corresponds to the one with threshold 799, having an expected cost $\mathbf{E}^{\delta,\alpha,\beta}[C(I,A)] = 5.19 \times 10^{-2}$. Finally, by using the optimal point for $ROC^{\alpha,\beta}$, as opposed to the original one, we get during operation an expected cost of $\mathbf{E}^{\text{operation}}[C(I,A)] = 2.73 \times 10^{-2}$. Therefore in this case, not only we have maintained our expected $5.19 \times 10^{-2} - security$ of the evaluation, but in addition the new optimal point actually performed better than the original one.

Notice that the evaluation of Figure 10 relates exactly to the problem we presented in the introduction, because it can be thought of as the evaluation of two IDSs. One IDS having a buffer threshold of length 399 and another IDS having a buffer threshold of length 773. Under ideal conditions we choose the IDS of buffer threshold length of 399 since it has a lower expected cost. However after evaluating the worst possible behavior of the IDSs we decide to select the one with buffer threshold length of 773.

An alternative view can be achieved by the use of IDOC curves. In Figure 11(a) we see the original IDOC curve during the evaluation period. These curves give a false sense of confidence in the IDS. Therefore we study the IDOC curves based on $ROC^{\alpha,\beta}$ in Figure 11(b). In Figure 11(c) we can see how the IDOC of the actual operating environment follows more closely the IDOC based on $ROC^{\alpha,\beta}$ than the original one.



(a) Original IDOC obtained during the evaluation period

(b) IDOC obtained from $ROC^{\alpha,\beta}$

(c) IDOC during operation time

Figure 11. Robust IDOC evaluation

6. Conclusions and Future Work

There are two main problems that any empirical test of an IDS will face. The first problem relates to the inferences that once can make about any IDS system based on experiments alone. An example is the low confidence on the estimate for the probability of detection in the ROC. A typical way to improve this estimate in other classification tasks is through the use of error bars in the ROC. However, since tests of IDSs include very few attacks and their variations, there is not enough data to provide an accurate significance level for the bars. Furthermore, the use of error bars and any other cross-validation technique gives the average performance of the classifier. However, this brings us to the second problem, and it is the fact that since the IDSs are subject to an adversarial environment, evaluating an IDS based on its average performance is not enough. Our intruder model tries to address these two problems, since it provides a principled approach to give us the worst case performance of a detector.

The extent by which the analysis with a $(\delta, \alpha, \beta) - intruder$ will follow the real operation of the IDS will depend on how accurately the person doing the evaluation of the IDS understands the IDS and its environment, for example, to what extent can the IDS be evaded, how well the signatures are written (e.g. how likely is it that normal events fire alarms) etc. However, by assuming *robust* parameters we are actually assuming a pessimistic setting, and if this pessimistic scenario never happens, we might be operating at a suboptimal point (i.e. we might have been too pessimistic in the evaluation).

Finally we note that IDOC curves are a general method not only applicable to IDSs but to any classification algorithm whose classes are heavily imbalanced (very small or very large p). We plan to propose their use in other fields as a general alternative to ROCs for these type of classification problems. In particular, we point out that another choice for the x-axis on an IDOC curve is to select $1 - \Pr[I = 1|A = 1] = \Pr[I = 0|A = 1]$, instead of $\Pr[I = 1|A = 1]$. This can be done in order to mimic the ROC evaluation, since $\Pr[I = 0|A = 1]$ intuitively represents the *Bayesian false alarm rate*. That is, the x-axis would then represent the probability that the IDS operator will not find an intrusion when he investigates an alarm report (informally, it would represent the waste of time for the operator of the IDS). The final decision on which x-axis to use will depend on the personal interpretation of the user.

References

[1] The MIT lincoln labs evaluation data set, DARPA intrusion detection evaluation. Available at http://www.ll.mit.edu/IST/ideval/index.html.

[2] Software for empirical evaluation of IDSs. Available at http://www.cshcn.umd.edu/research/IDSanalyzer.

[3] S. Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS '99)*, pages 1–7, November 1999.

[4] S. Buchegger and J.-Y. Le Boudec. Nodes bearing grudges: Towards routing security, fairness, and robustness in mobile ad hoc networks. In *Proceedings of Tenth Euromicro PDP (Parallel, Distributed and Network-based Processing)*, pages 403 – 410, Gran Canaria, January 2002.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc, 1991.

[6] G. Di Crescenzo, A. Ghosh, and R. Talpade. Towards a theory of intrusion detection. In *ESORICS 2005, 10th European Symposium on Research in Computer Security*, pages 267–286, Milan, Italy, September 12-14 2005. Lecture Notes in Computer Science 3679 Springer.

[7] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In D. Barbara and S. Jajodia, editors, *Data Mining for Security Applications*. Kluwer, 2002.

[8] S. Forrest, S. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for unix processes. In *Proceedigns of the 1996 IEEE Symposium on Security & Privacy*, pages 120–12, Oakland, CA, USA, 1996. IEEE Computer Society Press.

[9] J. E. Gaffney and J. W. Ulvila. Evaluation of intrusion detectors: A decision theory approach. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pages 50–61, Oakland, CA, USA, 2001.

[10] G. Gu, P. Fogla, D. Dagon, W. Lee, and B. Skoric. Measuring intrusion detection capability: An information-theoretic approach. In *Proceedings of ACM Symposium on InformAtion, Computer and Communications Security (ASIACCS '06)*, Taipei, Taiwan, March 2006.

[11] H. Handley, C. Kreibich, and V. Paxson. Network intrusion detection: Evasion, traffic normalization, and end-to-end protocol semantics. In *10th USENIX Security Symposium*, 2001.

[12] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In *IEEE Symposium on Security & Privacy*, pages 211–225, Oakland, CA, USA, 2004.

[13] C. Kruegel, E. Kirda, D. Mutz, W. Robertson, and G. Vigna. Automating mimicry attacks using static binary analysis. In *Proceedings of the 2005 USENIX Security Symposium*, pages 161–176, Baltimore, MD, August 2005.

[14] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur. Bayesian event classification for intrusion detection. In *Proceedings of the 19th Annual Computer Security Applications Conference (ACSAC)*, pages 14–24, December 2003.

[15] C. Kruegel, D. Mutz, W. Robertson, G. Vigna, and R. Kemmerer. Reverse Engineering of Network Signatures. In *Proceedings of the AusCERT Asia Pacific Information Technology Security Conference*, Gold Coast, Australia, May 2005.

[16] W. Lee and S. J. Stolfo. Data mining approaches for intrusion detection. In *Proceedings of the 7th USENIX Security Symposium*, 1998.

[17] W. Lee, S. J. Stolfo, and K. Mok. A data mining framework for building intrusion detection models. In *Proceedings of the IEEE Symposium on Security & Privacy*, pages 120–132, Oakland, CA, USA, 1999.

[18] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition*, volume 2, pages 12–26, January 2000.

[19] D. J. Marchette. A statistical method for profiling network traffic. In *USENIX Workshop on Intrusion Detection and Network Monitoring*, pages 119–128, 1999.

[20] S. Marti, T. J. Giuli, K. Lai, and M. Baker. Mitigating routing misbehavior in mobile ad hoc networks. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 255–265. ACM Press, 2000.

[21] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pages 1895–1898, Rhodes, Greece, 1997.

[22] J. McHugh. Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by the Lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC)*, 3(4):262–294, November 2000.

[23] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 2nd edition edition, 1988.

[24] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, March 2001.

[25] T. H. Ptacek and T. N. Newsham. Insertion, evasion and denial of service: Eluding network intrusion detection. Technical report, Secure Networks, Inc., January 1998.

[26] M. Schonlau, W. DuMouchel, W.-H. Ju, A. F. Karr, M. Theus, and Y. Vardi. Computer intrusion: Detecting masquerades. Technical Report 95, National Institute of Statistical Sciences, 1999.

[27] U. Shankar and V. Paxson. Active mapping: Resisting NIDS evasion without altering traffic. In *Proceedings of the 2003 IEEE Symposium on Security & Privacy*, pages 44–61, Oakland, CA, USA, 2003.

[28] S. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan. Cost-based modeling for fraud and intrusion detection: Results from the JAM project. In *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition*, pages 130–144, January 2000.

[29] K. Tan, K. Killourchy, and R. Maxion. Undermining an anomaly-based intrusion detection system using common exploits. In *Proceeedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID 2002)*, pages 54–73, Zurich, Switzerland, October 2002.

[30] K. Tan, J. McHugh, and K. Killourhy. Hiding intrusions: From the abnormal to the normal and beyond. In *Information Hiding: 5th International Workshop*, pages 1–17, Noordwijkerhout, The Netherlands, October 2002.

[31] H. L. Van Trees. *Detection, Estimation and Modulation Theory, Part I*. Wiley, New York, 1968.

[32] G. Vigna, S. Gwalani, K. Srinivasan, E. Belding-Royer, and R. Kemmerer. An Intrusion Detection Tool for AODV-based Ad Hoc Wireless Networks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, pages 16–27, Tucson, AZ, December 2004.

[33] G. Vigna, W. Robertson, and D. Balzarotti. Testing Network-based Intrusion Detection Signatures Using Mutant Exploits. In *Proceedings of the ACM Conference on Computer and Communication Security (ACM CCS)*, pages 21–30, Washington, DC, October 2004.

[34] D. Wagner and P. Soto. Mimicry attacks on host-based intrusion detection systems. In *Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS)*, pages 255–264, Washington D.C., USA, 2002.

[35] C. Warrender, S. Forrest, and B. Pearlmutter. Detecting intrusions using system calls: Alternative data models. In *Proceedings of the 1999 IEEE Symposium on Security & Privacy*, pages 133–145, Oakland, CA, USA, May 1999.

[36] Y. Zhang, W. Lee, and Y. Huang. Intrusion detection techniques for mobile wireless networks. *ACM/Kluwer Mobile Networks and Applications (MONET)*, 9(5):545–556, September 2003.